Chapter 3

Materials and Methods

3.1 DATABASES DESIGN AND IMPLEMENTATION

The databases SPGDB, HIGDB, MDB and EDB have been created and implemented on Solaris server and the server is extraordinarily power-driven by 2.66 gigahertz processor with principle memory of about 4 GB FDIMM. The Solaris server was selected to host these databases because of its features like versatility, capability to handle large queries and it is more secured server. The complete data of SPGDB, HIGDB, MDB and EDB were stored and managed in a relational database management system, My Structured Query Language (MySQL). A schema was built in MySQL under which 5 tables were created for each database. The search engine implemented in four databases was written using PERL/CGI and PERL/DBI modules. The front-end input data part of four databases was encoded in HTML, JavaScript and Ajax which permits the web-forms to be user-friendly.

3.2 RETRIEVAL OF DATA

The retrieval methods of various data employed in SPGDB, HIGDB, MDB and EDB and the tools deployed in these databases are shown in Fig. 3.1.

3.2.1 NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION (NCBI) GENOME DATABASE

The genome is a complete set of DNA comprising a chromosome. The genome of an organism includes transcript sets and protein sequences translated from their corresponding nucleotide sequences. NCBI made an effort to provide the complete genome of various organisms at a single point to the researchers. The genomes of organisms available at NCBI are very well collaborated with GenBank and Reference Sequence (RefSeq). GenBank is a unit of international nucleotide sequence database collaboration and gathers annotated nucleotide sequences from researchers when they are publishing those sequences to a research journal. RefSeq
gives established note for genome annotation and various bioinformatics studies (Wheeler et al., 2007).

**Fig. 3.1:** Schematic representation of retrieval of various information and tools deployed in developed databases

The complete genomic information of *S. pneumoniae* strains, *H. influenzae* strains, *B. pseudomallei* strains and *Ebolavirus* species were retrieved from their corresponding Genome assembly and annotation report available in NCBI database. This report had information such as bioproject details, assembly, accession number, size and GC content of genome, number of bp and total number of genes and proteins present in that genome. All these information were exported and stored in “strain_list” table present in their respective schema in MySQL. The genome of all available strains of *S. pneumoniae*, *H. influenzae*, *B. pseudomallei* and *Ebolavirus* were downloaded from File Transfer Protocol (FTP) site of NCBI genome database in GenBank (gbk) format. An example for gbk format of genome is shown in Fig. 3.2. The figure is screenshot of part of a gbk file of *S. pneumoniae* R6 genome.
Contd...
Contd...
..

ORIGIN

```
1 ttgaagaaaa acaaatcttg gaactcgata ttgaaatttg cacaagaaag acctgactcga
61 tcagcagtct gacatctacg tattcatctc gcgacttcat acagacgtgc cc
121 actataatttc tataaggtgat atgggtggtg aaaaacta ctctaagat tt
181 attgtagtag cagttttcag aatatttagc gatggaataa ccctctacta tattttccc
241 aacactcaag atacagtact ccacagatct gaagaactca caatatttaac ttctctgagc
301 tattgctcaa aagttgagtt tattcttcac atatatgtc gcgttataggat cattacacaac
361 ttgataact ttttaagag gatggaataa gttggtgtgt ttacagtgtcc ttataaggctc
421 ttgaaagatt tggctctgac ataaacccct ctatttatct ctggaagcc gcggcctgtt
481 aagactacat tataaacgct tattgaaat gaacttcatca aatatcttcc tactgcgtgct
541 gtagaataata tcctctgacg aagctttata aatgccttct ttgctctcct aagctgtggc
601 gaagaagaga aagttgagtt tattcttcac atatatgtc gcgttataggat cattacacaac
661 cagtcctcga ggcgaagaaa agtggaaact caggaagat ttctcttcc cttaagcggc
721 ctctctgac caagaaaaaa gatgtgtcttc aagagtgatc gtagtccaa actcttggaa
781 ggaccctcag agagctcttg ccgctttgct atggaaggtg tgcacaaagc ccttccgaccc
841 cctgacctctg aacacagttc tcgacatatt tctaaatgaa cgaaaccttt ggtctcaaat
901 ttcgcaattgt gcacacttgtg ggcgacatgtg atccaaatgt tcgagctctt
961 gaggagcctg tcaacagcact ctatcttaat tcgctaagtt aaaaaatccaaa ggtataacact
1021 tattagatgg ctcgcaacac gcacaaaaag agtggagaaa atgttgccat cattgtgctc
1081 attcggattt tctgtggagt gacatgctga tctacaaattt tctacatgtt tcatgcaaat
1141 atggaagcag tgaagctgct tcctaatattt gttgttcgct gtaagagacg catgtatatta
1201 tctgagaagac tataactaca tggctttttc aaaaaagcag gatgttgggg ggggaaagat
1261 cattacagct tcagcttgat gcctggcaaa ataaaatgtc tgtatgcata gagaagataa
1321 ttacagctgtaaatagatc atcggaaaaa aaaaatcctt atatatgtgtaaacccttttacatggtttt
1381 ttattatttt tttttttttt tttttttttt acaagctcaaa aacatgccca gagactgctt
1441 aaagctcttt ttcgcaagat ttcgaacact ctattttcact tattttctttaataggct
1501 aatattatata agaggaatoc atgaccttt ttcacattaa taaaatattttttt ttcacacag
```
Fig. 3.2: Screenshot of part of a gbk file of *S. pneumoniae* R6 genome
From the gbk file of all strains, the details such as gene and protein names, locus tag, start and end position, direction, nucleotide sequences and their corresponding translated sequences were extracted for each feature (CDS, tRNA, rRNA, ncRNA and tmRNA) using PERL program. The extracted information were exported and stored in “gene_list” table in their respective schema in MySQL.

3.2.2 UNIVERSAL PROTEIN RESOURCE (UNIPROT)

UniProt is an accumulation of data on proteins. UniProt is a product of UniProt Consortium and this consortium is teamwork of three groups namely

- Protein information resource
- European bioinformatics institute
- Swiss institute of bioinformatics

UniProt databases are grouped into three categories: a) UniProt Knowledgebase (UniProtKB) b) UniProt Archive (UniParc) c) UniProt reference clusters (UniRef). UniProtKB is further subdivided into UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot is a collection of reviewed and physically curated entries and on the other hand, UniProtKB/TrEMBL is a collection of unreviewed and computerized entries. UniParc is an exhaustive and non-repetitious resource that consists of vast majority of openly accessible protein sequences throughout the world. It avoids the repeating of a protein sequence and therefore, a protein sequence will be exist only once in UniParc database. UniRef is a resource that gives a set of sequences which are grouped into clusters based on their sequence identity scores (Bairoch et al., 2005).

The entry name, status of the protein whether it is reviewed or unreviewed, protein and their gene names, organism name along with strain detail from which it is produced, gene ontology information, their sub cellular location, protein sequence and sequence length for each entry were downloaded from UniProt for all strains/species of *S. pneumoniae, H. influenzae, B. pseudomallei* and *Ebola virus*. The entries from both UniProtKB/Swiss-Prot and UniProtKB/TrEMBL database were considered. This
information were exported and stored in a table “protein_list” in their respective schema in MySQL.

3.2.3 PUBMED DATABASE

PubMed is a literature database which is created and controlled by NCBI team. The exploring option provided in PubMed Database has the ability to retrieve the information from MEDLINE database and provide the summaries on life sciences and biomedical subjects. MEDLINE is the principle element of PubMed. The important role of MEDLINE is to act as a vital asset for biological scientists and various journal clubs throughout the world (Masic and Milinovic, 2012). It covers almost all the subjects and areas. PubMed resource has more than twenty one million entries from over five thousand publications as well as online books.

The literature information on *S. pneumoniae, H. influenzae, B. pseudomallei* and *Ebolavirus* were searched in PubMed database individually. The Boolean operators were used for complex search. The search results were downloaded in summary format and using the PERL program, title and author of the publication, journal name in which the article is published, year, issue and volume of the publication and their corresponding PubMed identifier were extracted. The extracted details were then exported to a table “pubmed_list” in their respective schema in MySQL.

3.2.4 PROTEIN DATA BANK (PDB)

PDB is the structural database consisting of 3-D structures of biological macromolecules which is maintained by a consortium named research collaboratory for structural bioinformatics. The database contains the data evolved from X-ray crystallography and Nuclear Magnetic Resonance spectroscopy which were deposited in PDB by structural biologist from all over the world. Most of the 3-D structures of proteins were from X-ray crystallography method. These structures can be freely accessible by all the users. PDB plays an important role in the field of structural biology and structural genomics (Berman et al., 2002). There are more than 1, 14,600
3-D structures of biological macromolecules available in PDB at the end of the year 2015.

The 3-D structures of proteins obtained from S. pneumoniae, H. influenzae, B. pseudomallei and Ebolavirus of all strains/species were downloaded from PDB in .pdb format. Using the PERL program, the title of the structure, protein name and PDB identifier, source of the organism and their corresponding UniProt identifier were extracted and exported to a table “pdb_list” in their respective schema in MySQL. The schema and list of tables created in MySQL for SPGDB, HIGDB, MDB and EDB are shown in Table 3.1.

Table 3.1:

<table>
<thead>
<tr>
<th>Database</th>
<th>Schema Name</th>
<th>Table name for Strain/species information</th>
<th>Table name for Gene information</th>
<th>Table name for Protein information</th>
<th>Table name for Structural information</th>
<th>Table name for Literature information</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPGDB</td>
<td>SPGDB</td>
<td>strain_list</td>
<td>gene_list</td>
<td>protein_list</td>
<td>pdb_list</td>
<td>pubmed_list</td>
</tr>
<tr>
<td>HIGDB</td>
<td>HIGDB</td>
<td>strain_list</td>
<td>gene_list</td>
<td>protein_list</td>
<td>pdb_list</td>
<td>pubmed_list</td>
</tr>
<tr>
<td>MDB</td>
<td>MDB</td>
<td>strain_list</td>
<td>gene_list</td>
<td>protein_list</td>
<td>pdb_list</td>
<td>pubmed_list</td>
</tr>
<tr>
<td>EDB</td>
<td>EDB</td>
<td>organism_list</td>
<td>gene_list</td>
<td>protein_list</td>
<td>pdb_list</td>
<td>pubmed_list</td>
</tr>
</tbody>
</table>

3.2.5 GENOME ATLAS DATABASE

The number of bacterial genomes being sequenced has been increased recently. Hence, there is a requirement for an adaptable and a simply maintained structure to save the sequence information and results from various in-silico analyses. The comparative study results on different characteristics of phylogenetically similar organisms are not easily accessible to numerous scientists. In view of above, the genome atlas database was developed which is an integrative database. The search engine developed in genome atlas database makes possible for the users to search within a particular kingdom. In genome atlas database, the genome map can be downloaded either in vector graphic format or portable network graphics format (Hallin and Ussery, 2004) and the genome maps may be linear or circular based on
the bacterial chromosome structure. The genome maps available in genome atlas database provide an information on intrinsic curvature, stacking energy, position preference, various annotations (CDS+, CDS-, fRNA and tRNA), global direct and inverted repeats, guanine (G) and cytosine (C) skew and percentage of adenine and thymine content of bacterial chromosomes. The genome map of all available strains of *S. pneumoniae*, *H. influenzae* and *B. pseudomallei* were downloaded and moved to the image folder present in the respective database directory in the Solaris server.

The data retrieved from various resources were cross verified with each other and if there is any discrepancy of data, the duplicates were removed when they were exported to MySQL.

3.3 IMPLEMENTATION OF TOOLS

The following tools have been implemented in SPGDB, HIGDB, MDB and EDB databases

- Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990)
- Generic Genome Browser (GBrowse) (Stein et al., 2002)
- JmolApplet

3.3.1 BLAST

The BLAST is a sequence similarity search tool which is used to recognize the biological function of putative sequences and to locate the similarity regions locally between a pair of sequence. The tool compares nucleotide or amino acid sequences with the sequences present in different resources and measures the statistical significance of matched regions. The results obtained from BLAST can be helpful in deducing the functional and phylogenetic relationships among the pair of sequences and also aid in the identification of gene family to which the sequence belong. In spite of BLAST’s adaptability and compliancy to numerical calculations, BLAST provides the results more quickly than any other comparative tools available and the BLAST has high sensitivity.
3.3.1.1 BLAST ALGORITHM

BLAST employs heuristics methods to generate outputs rapidly, which in fact shows that it makes the use of shortcut to produce the results quickly. It is noteworthy to cut the BLAST search into three steps:

1) Setup
2) Preliminary search
3) Traceback

In the setup step, BLAST examines the query sequence, the parameters which are selected and the database against which the search is perform. Initially, BLAST verifies the query sequence for minimal complicated regions or repeated regions. Then, it forms a small group of words which is of steady length depending on the query sequence. These small groups of words are used to start searching for matches in the database sequences. The next step after setting up of word for BLAST search is the preliminary search where more steps are carried out on each and every sequence present in the public database. BLAST will search for the word to match in the database sequence and once the word gets matched with the subject sequence, it begins ungapped extension. After the ungapped extension, it obtains a score which is used to continue the search with gapped extension and this gapped extension gets a score. The final score values are saved and the sequence with low score are eliminated. In the last step, the insertions and deletions of gaps are measured and at lastly, display the sequence that has good score (Altschul et al., 1990).

3.3.1.2 STEPS INVOLVED IN THE IMPLEMENTATION OF BLAST

Step 1: Download the standalone BLAST executables from FTP site of NCBI

Step 2: Installation of BLAST

Step 3: Set the environment variables

Step 4: Create a directory for BLAST database

Step 5: Set the variables for BLAST database
Step 6: Download the FASTA formatted nucleotide and protein sequences for each strain

Step 7: Format the downloaded FASTA sequences into BLAST database using makeblastdb program

Step 8: Move all files to the BLAST database of NCBI bin folder

3.3.2 GBROWSE

GBrowse is an elementary online based genome browser and the configuration of browser is very greatly built. The browser is a mixture of database and user-friendly web pages and it is used for controlling and showing the complete details of genome graphically. GBrowse is developed by Generic Model Organism Database (GMOD) project and GBrowse is the most important part of GMOD project. The application is implemented by using PERL language. The browser has been effectively implemented and employed in various model organism specific resources like WormBase, FlyBase and PlasmoDB (Stein et al., 2002).

3.3.2.1 STEPS INVOLVED IN THE IMPLEMENTATION OF GBROWSE IN DATABASES

Step 1: Download the GBrowse executables from GMOD

Step 2: Install GBrowse along with its dependencies

Step 3: Download the genome sequence of each strain/species in genome feature format and FASTA format from FTP site of NCBI

Step 4: Move all downloaded genome files into their respective database directory of GBrowse which is present in /lib

Step 5: Prepare the configuration file for each strain/species and moved to their respective GBrowse directory which is present in /etc
3.3.3 JMOLAPPLET

JmolApplet is an intuitive web program which can be incorporated into websites or databases to show 3-D structure of the biological macromolecules or chemical structures in different ways. The application is developed based on the JAVA platform. It works better and faster than the Chime plug-in and it is the principle web program for the online approachable chemical databases. JmolApplet accepts extensive variety of file formats like PDB format.

3.3.3.1 STEPS INVOLVED IN THE IMPLEMENTATION OF JMOLAPPLET IN DATABASES

**Step 1:** Download the JmolApplet on Solaris server along with the recent version of JAVA applet

**Step 2:** Move the downloaded PDB structural coordinates to their respective database present in Jmol directory