

DEVELOPMENT OF A KNOWLEDGEBASE FOR VARIOUS CANCER TYPES THROUGH SYSTEM LEVEL ANALYSIS

Satyam Kapoor (121502)

Tuhina Srivastava (121508)

Sheena Singh (121516)

Shubham Thakur (121520)

Under supervision of Dr. Tiratha Raj Singh



June – 2016

Submitted in partial fulfilment of the Degree of

Bachelor of Technology

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT

TABLE OF CONTENTS

Chapter No.	Chapter	Page No.
	Certificate	i
	Declaration	ii
	Acknowledgement	iii
	Abstract	iv
	Table of figures	v
1.	Introduction	1
2.	Ideal Approach	3
3.	Data	7
4.	Sources Of Data	16
5.	Technology Stack	17
6.	Web Design	22
7.	Challenges	31
	Web References	32
	Reference Articles	33

CERTIFICATE

This is to certify that the work titled “**DEVELOPMENT OF A KNOWLEDGEBASE FOR VARIOUS CANCER TYPES THROUGH SYSTEM LEVEL ANALYSIS**” submitted by “**SATYAM KAPOOR, TUHINA SRIVASTAVA, SHEENA SINGH, SHUBHAM THAKUR**” in partial fulfilment for the award of degree of B. Tech of Jaypee University of Information Technology, Wanknaghat has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of the Supervisor

Name of the Supervisor

Designation

Date

DECLARATION

I hereby declare that the work presented in this report entitled “Development of a Knowledgebase for various cancer types through System level analysis” in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Bioinformatics submitted in the Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Wagnaghat, Solan-173234, Himachal Pradesh is an authentic record of my own work carried out over a period from August 2015 to June 2016 under the supervision of Dr Tiratha Raj Singh, Assistant Professor (Senior Grade), Department of Biotechnology and Bioinformatics.

The matter embodied in the report has not been submitted for the award of any other degree or diploma

Satyam Kapoor (121502)

Tuhina Srivastava (121508)

Sheena Singh (121516)

Shubham Thakur (121520)

ACKNOWLEDGEMENT

We owe a great thanks to many people who have been helping and supporting us during this project. Our deepest thanks to our guide Dr. Tiratha Raj Singh for guiding and correcting us at every step of our work with attention and care. He has taken a lot of pain to go through the project and make necessary corrections as and when needed. We would also like to thank our institution and the Department of Biotechnology and Bioinformatics without whom this project would have been a distant reality. We also extend our heartfelt thanks to our family and well-wishers.

Satyam Kapoor (121502)

Tuhina Srivastava (121508)

Sheena Singh (121516)

Shubham Thakur (121520)

ABSTRACT

The two most important components of our project are the types of cancers and a combination of databases which is known as a knowledgebase. A database contains collection of information that is organised so that it can easily be accessed, managed and updated. We have studied various cancer databases and compiled the most common and important attributes of cancer types. Our online database is known as Cancertome and aims to provide users with the data, the users can narrow down their search by our various specific search options.

Satyam Kapoor
Tuhina Srivastava	Dr. Tiratha Raj Singh
Sheena Singh	Date:
Shubham Thakur	

TABLE OF FIGURES

Fig	Figure Name	Page no.
1	Homepage	1
2	Homepage	2
3	Pie chart representing statistics of common cancer types	3
4	Statistics representing cancer affected males and females	3
5	Statistics of number of death due to various types of cancers	4
6	Worldwide distribution of Cancers and p53 Mutations	12
7	Overall backend process	19
8	OLTP and OLAP	20
9	PHP framework popularity	21
10	Z Pattern	22
11	Homepage of Cancertome	24
12	Search Options	24, 25
13	Sample Result Page	26
14	Mitochondrial Search Page	26
15	Download Page	27
16	Contact Page	27
17	phpMyAdmin Database	29
18	Sample Table Structure	29
19	Sample Table Data	30

CHAPTER 1 INTRODUCTION

1.1 What is a knowledge base?

A **knowledge base** is not a static collection of information, but a dynamic resource that may itself have the capacity to learn. It typically can help link and integrate all available knowledge sources. It includes explicit knowledge (various kinds of databases stored in existing information systems) and inexplicit knowledge (practical experience, skills, thought and thinking method in the brain of the experts / people). It is a technology used to store complex structured and unstructured information used by a computer system. In general, a knowledge base is a centralized repository for information: a public library, a database of related information about a particular subject.

1.1.1 Advantages

- **Improve productivity** by implementing a tool that employees can use to find the information they need, when they need it. According to a study, employees spend 20% of their time at work searching for the information they need to get work done. That's one full day per week!
- **Encourage collaboration** with a tool that multiple people can use to work on content together. Allowing them to collaborate on a piece of content.
- **Reserve email for one-to-one communication** by encouraging employees to utilize the knowledge base for sharing information that might be helpful to a larger group, or to future employees. This eliminates any versioning issues, reduces confusion, and everyone can access information easily without having to search in email.
- **Ease of Access** A knowledge base can give customers easy access to information that would otherwise require cumbersome methods to gather information from different databases.

1.2 CANCERTOME (www.cancertome.online)



Fig 1. Homepage

As the title of our project suggests, we have created a knowledgebase **CANCERTOME** an online tool which can be accessed via url **www.cancertome.online**.

It is a repository of all the cancers we have worked upon. Important information can be retrieved regarding any specific cancer. We will be able to analyze and reuse information. Parameters can be set accordingly to filter our search.

Important attributes for cancers considered are mentioned further in the section 3.2. Our knowledgebase **CANCERTOME** serves as a repository for the information on cancer you want to search upon.

The data has been collected from various sources which are mentioned in the section 4.



Fig 2. Homepage

CHAPTER 2 IDEAL APPROACH

We start with minimal data in our project and step by step proceed to our knowledgebase. We selected the cancers which are most commonly affecting the masses and take up some of the attributes regarding them. Building up a knowledgebase is actually a mixture of gathering the data and applying some of the data mining techniques plus a good level of web development skills. Our team is working through all the aspects of this plan.

2.1. Identify

- Build an approach so as to achieve your goal.
- Specify and identify the characteristics to be associated with your cancers.

We chose the cancers on the basis of how commonly they affect the masses. Various statistical data were considered for this area

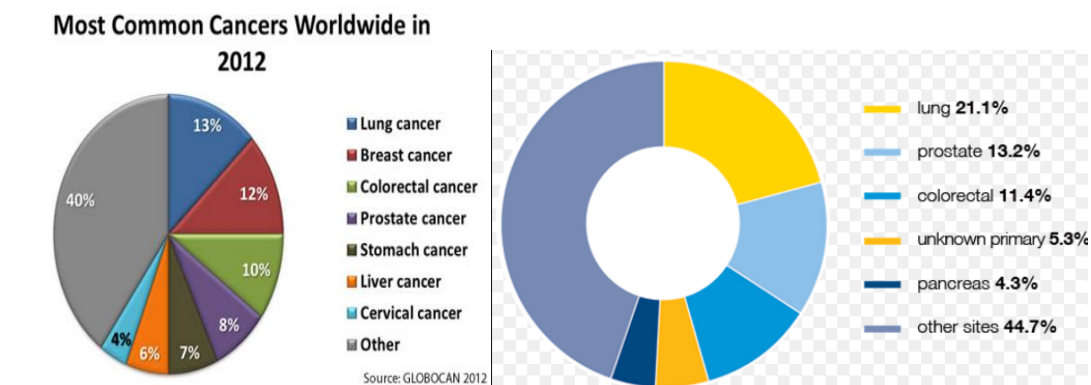


Fig 3: Pie chart representing statistics of common cancer types



Fig 4: Statistics representing cancer affecting males and females

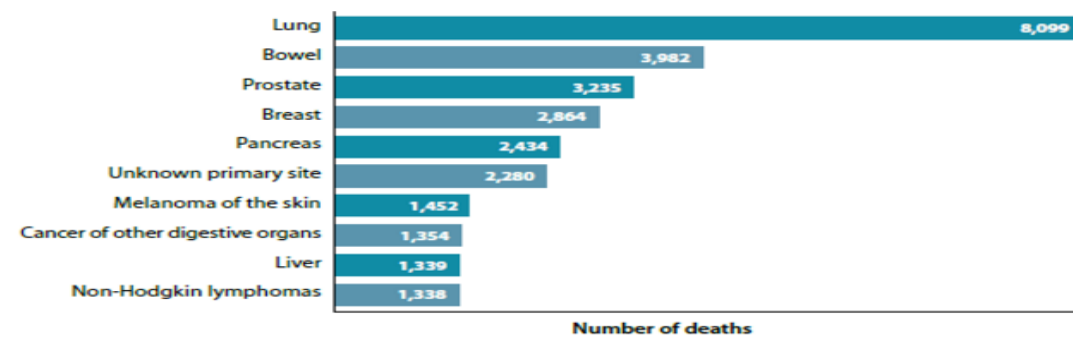


Fig 5: Statistics of number of deaths due to various types of cancer

Data source: GLOBOCAN 2012: Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012

Based on this approach we deduced 7 cancers that we would be working upon.

- Brain cancer
- Breast cancer
- Endometrial cancer
- Lung – small cell lung cancer, non small cell lung cancer
- Pancreatic cancer
- Colorectal cancer.

2.2. Data Collection

- All the information regarding the cancers was collected from the already existing databases.
- The attributes were kept in mind and then the data was extensively searched upon. Various pre existing databases were searched for these attributes.
- We also had to keep in mind if our data was downloadable or not.

The attributes we considered were:

- General information
- Morphology
- Single Nucleotide Polymorphism(SNP)
- p53 Mutations
- MiRNA
- Long non coding RNA

- Gene ontology (GO)
- Differential expression
- Tumour expression
- Mitochondria role in Cancer

Further information about these are provided in the section 3.2.

2.3. Data Refinement

- In this step we first performed refinement of data.
- The data which is as per our requirement of attributes or cancer was retrieved.
- All the selected cancers and their respective attributes were retained and the rest were eliminated.

We performed this work using PhpMyAdmin via SQL Queries.

- Eg. In order to retain only the lung cancer fields the query used was:

```
DELETE FROM TABLE1
WHERE CANCER_TYPE!="LUNG CANCER" .
```

- The next step was to remove data redundancy. Data redundancy is defined as the storing of the same data in multiple locations.

Data redundancy was removed with the following SQL query-

```
SELECT DISTINCT 'attribute'
FROM 'TABLE1'
```

2.4 DATA MODEL CONSTRUCTION:

Next step is to create a data model i.e. Entity Relationship Model.

- An entity–relationship model (ER model) is a data model for describing the data or information in an abstract way that lends itself to ultimately being implemented in a database such as a relational database.
- The main components of ER models are entities (things) and the relationships that can exist among them. The data is represented as components (entities) that are linked with each other by relationships that express the dependencies and requirements between them.

- An ER model is typically implemented as a database. In the case of a relational database, which stores data in tables, every row of each table represents one instance of an entity.

2.5. WEB DESIGN

Web designing is an integral part of building this knowledgebase.

- Once the whole process of data collection and data model has been implied we now shift upon web designing.

We need to provide users with a user friendly and interactive interface which provides easy access of information.

- The information retrieved is to be displayed in clear tabular formats and also a well defined GUI will allow easy navigation through the website.
- Also on server side a systematic database would be easy to handle and allow efficient up-gradation and change of additional data.

3.1 Types of cancers included in our Knowledgebase

3.1.1. Brain Cancer

- A brain tumor or intracranial neoplasm occurs when abnormal cells form within the brain.
- The most common primary brain tumors are usually named for the brain tissue type (including brain stem cancers) from which they originally developed.

These are:

- Gliomas
- Meningiomas
- Pituitary adenomas
- Vestibular schwannomas and
- Primitive neuroectodermal tumors (medulloblastomas).

All types of brain tumors may produce symptoms that vary depending on the part of the brain involved.

- These may include headaches, seizures, problem with vision, vomiting, and mental changes.
- The prognosis (chance of recovery) depends on many factors, including age, tumor size, tumor type, and where the tumor is in the CNS.

3.1.2. Breast Cancer

- Breast cancer is a malignant tumor arising from the cells of the breast.
- Although breast cancer predominantly occurs in women, it can also affect men.
- Breast cancer symptoms and signs include a lump in the breast or armpit, bloody nipple discharge, inverted nipple, orange-peel texture or dimpling of the breast's skin, breast pain or sore nipple, swollen lymph nodes in the neck or armpit, and a change in the size or shape of the breast or nipple.

- Risk factors for breast cancer include age, family history, personal history, breast tissue, race etc.
- Breast cancer is diagnosed during a physical exam, by self-examination of the breasts, mammography, ultrasound testing, and biopsy.
- Treatment of breast cancer depends on the type of cancer and its stage (0-IV) and may involve surgery, radiation, or chemotherapy.

3.1.3. Endometrial Cancer

- Endometrial cancer is a cancer that arises from the endometrium (the lining of the uterus or womb).
- It is the result of the abnormal growth of cells that have the ability to invade or spread to other parts of the body.
- The first sign is most often vaginal bleeding not associated with a menstrual period. Other symptoms include pain with urination or sexual intercourse, or pelvic pain. Endometrial cancer occurs most commonly after menopause.
- Approximately 75% of women with endometrial cancer are postmenopausal. For the 25% of endometrial cancers in patients who are perimenopausal or premenopausal, the symptoms suggestive of cancer may be more subtle.
- The normal menstrual bleeding pattern during this period should become lighter and lighter and further and further apart. Heavy, frequent menstrual periods or intermenstrual bleeding must be evaluated.

3.1.4 Colorectal Cancer

- It is the development of cancer from the colon or rectum (parts of the large intestine).
- It is due to the abnormal growth of cells that have the ability to invade or spread to other parts of the body.
- Some risk factors include diet, obesity, smoking, and lack of physical activity. Dietary factors that increase the risk include red and processed meat as well as alcohol.
- Another risk factor is inflammatory bowel disease.

- Some of the inherited genetic disorders that can cause colorectal cancer include familial adenomatous polyposis and hereditary non-polyposis colon cancer, however, these represent less than 5% of cases.
- It typically starts as a benign tumor, often in the form of a polyp, which over time becomes cancerous. Most colorectal cancers develop from polyps.
- Removal of colon polyps can prevent colorectal cancer. Colon polyps and early cancer may have no symptoms. Therefore regular screening is important.
- Treatment of colorectal cancer depends on the location, size, and extent of cancer spread, as well as the health of the patient.
- Surgery is the most common treatment for colorectal cancer. Chemotherapy can extend life and improve quality of life for those who have had or are living with metastatic colorectal cancer. It can also reduce the risk of recurrence in patients found to have high-risk colon cancer findings at surgery.

3.1.5 Pancreatic Cancer

- Pancreatic cancer arises when cells in the pancreas, a glandular organ behind the stomach, begin to multiply out of control and form a mass.
- These cancerous cells have the ability to invade other parts of the body.
- There are a number of types of pancreatic cancer. The most common, pancreatic adenocarcinoma, accounts for about 85% of cases. One to two in every hundred cases of pancreatic cancer are neuroendocrine tumors, which arise from the hormone-producing cells of the pancreas. These are generally less aggressive than pancreatic adenocarcinoma.
- Signs and symptoms of the most common form of pancreatic cancer may include yellow skin, abdominal or back pain, unexplained weight loss, light-colored stools, dark urine and loss of appetite.
- There are usually no symptoms in the disease's early stages, and symptoms that are specific enough to suggest pancreatic cancer typically do not develop until the disease has reached an advanced stage. By the time of diagnosis, pancreatic cancer has often spread to other parts of the body.
- Pancreatic cancer rarely occurs before the age of 40, and more than half of cases of pancreatic adenocarcinoma occur in those over 70.
- Risk factors for pancreatic cancer include tobacco smoking, obesity, diabetes, and certain rare genetic conditions.

3.1.6 Lung cancer

- Lung cancer is the number one cause of cancer deaths in both men and women worldwide.
- Cigarette smoking is the principal risk factor for development of lung cancer. Passive exposure to tobacco smoke also can cause lung cancer.
- The two types of lung cancer, which grow and spread differently, are small cell lung cancers (SCLC) and non-small cell lung cancers (NSCLC).
- About 10% to 15% of lung cancers are SCLC. NSCLC makes up about 80% to 85% of lung cancers. The 3 main types of NSCLC are adenocarcinoma, squamous cell carcinoma, and large cell carcinoma.
- The symptoms are coughing, coughing up blood, wheezing, or shortness of breath, weight loss, weakness, fever, or clubbing of the fingernails. Symptoms due to the cancer mass pressing on adjacent structures: chest pain, bone pain, superior vena cava obstruction, or difficulty swallowing
- If the cancer grows in the airways, it may obstruct airflow, causing breathing difficulties. The obstruction can lead to accumulation of secretions behind the blockage, and predispose to pneumonia.
- Many of the symptoms of lung cancer (poor appetite, weight loss, fever, fatigue) are not specific. In many people, the cancer has already spread beyond the original site by the time they have symptoms and seek medical attention.
- Symptoms that suggest the presence of metastatic disease include weight loss, bone pain and neurological symptoms (headaches, fainting, convulsions, or limb weakness).
- Common sites of spread include the brain, bone, adrenal glands, opposite lung, liver, pericardium, and kidneys. About 10% of people with lung cancer do not have symptoms at diagnosis; these cancers are incidentally found on routine chest radiography.

3.2 Main Attributes

3.2.1. General information

It contains the very general information like gene related to the cancers and studies carried out with respect to them.

3.2.2. Morphology

- The morphology of a cancer refers to the histological classification of the cancer tissue (histopathological type) and a description of the course of development that a tumour is likely to take: benign or malignant (behaviour).
- It will indicate where the cancer first started affecting the body.

3.2.3. Single Nucleotide Polymorphism (SNP)

- A single nucleotide polymorphism is a variation in a single nucleotide that occurs at a specific position in the genome, where each variation is present to some appreciable degree within a population (e.g. >1%).
- SNPs underlie differences in our susceptibility to disease; a wide range of human diseases.
- Genetic variation in the human genome is an emerging resource for studying cancer, a complex set of diseases characterised by both environmental and genetic contributions.
- The number of common germ-line variants is great, on the order of 10-15 million per person, and represents a remarkable opportunity to investigate the etiology, inter-individual differences in treatment response and outcomes of specific cancers.
- The study of genetic variation can elucidate critical determinants in environmental exposure and cancer, which could have future implications for preventive and early intervention strategies. However, we are in the initial stages of characterizing the tools (i.e., the single-nucleotide polymorphism, SNP) to rigorously analyze the genetic contributions to complex diseases, such as cancer.
- If the promise of the genomic era is to be realized, we must integrate this information into new strategies for implementation in both public health measures and, most importantly, provision of individual cancer-related care.

3.2.4. P53 Mutations

- Tumor protein p53, also known as p53, cellular tumor antigen p53 (UniProt name), phosphoprotein p53, tumor suppressor p53, antigen NY-CO-13, or transformation-related protein 53 (TRP53), is any isoform of a protein encoded by homologous genes in various organisms, such as TP53 (humans) and Trp53 (mice).
- This homolog is crucial in multicellular organisms, where it prevents cancer formation, thus, functions as a tumor suppressor.

- p53 has been described as "the guardian of the genome" because of its role in conserving stability by preventing genome mutation. Hence TP53 is classified as a tumor suppressor gene.
- It has become apparent that tumor-associated p53 mutations can provoke activities that are different to those resulting from simply loss of wild-type tumor-suppressing p53 function. Many of these mutant p53 proteins acquire oncogenic properties that enable them to promote invasion, metastasis, proliferation and cell survival. Here we highlight some of the emerging molecular mechanisms through which mutant p53 proteins can exert these oncogenic functions.
- The International Cancer Genome Consortium has established that the TP53 gene is the most frequently mutated gene (>50%) in human cancer, indicating that the TP53 gene plays a crucial role in preventing cancer formation. TP53 gene encodes proteins that bind to DNA and regulate gene expression to prevent mutations of the genome.

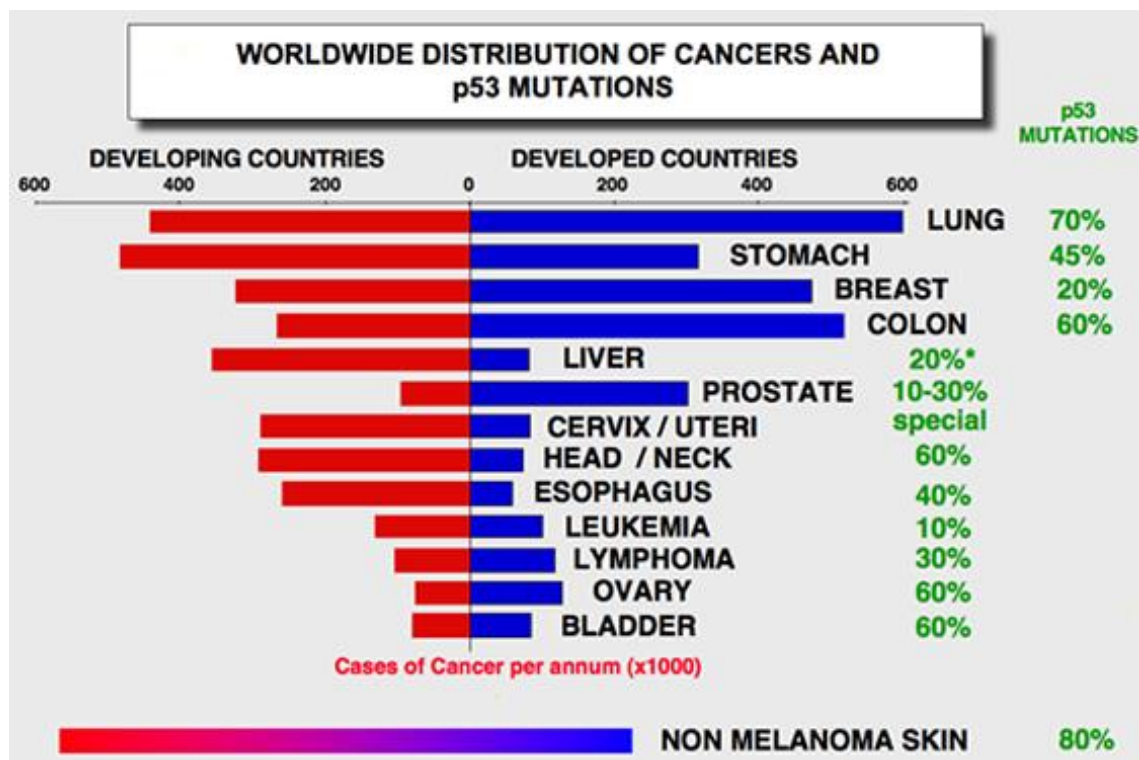


Fig 6: Worldwide distribution of cancers and p53 mutations

Data Source: The p53 Website (http://p53.free.fr/Database/p53_cancer_db.html)

3.2.5. miRNA

- A micro RNA (abbreviated miRNA) is a small non-coding RNA molecule (containing about 22 nucleotides) found in plants, animals, and some viruses, which functions in RNA silencing and post-transcriptional regulation of gene expression.
- MicroRNAs (miRNAs) are causing tremendous excitement in cancer research. MiRNA expression profiles demonstrate that many miRNAs are deregulated in human cancers. MiRNAs have been shown to regulate oncogenes, tumour suppressors and a number of cancer-related genes controlling cell cycle, apoptosis, cell migration and angiogenesis.
- MiRNAs encoded by the *mir-17-92 cluster* have oncogenic potential and others may act as tumour suppressors. Some miRNAs and their target sites were found to be mutated in cancer. MiRNAs may have great diagnostic potential for human cancer and even miRNA-based cancer therapies may be on the horizon.

3.2.6 Long non coding RNA

- Long non-coding RNAs (lncRNA) are non-protein coding transcripts longer than 200 nucleotides. This somewhat arbitrary limit distinguishes long ncRNAs from small regulatory RNAs such as microRNAs (miRNAs), short interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), and other short RNAs.
- Advances in RNA sequencing technologies have revealed the complexity of our genome. Non-coding RNAs make up the majority (98%) of the transcriptome, and several different classes of regulatory RNA with important functions are being discovered. Understanding the significance of this RNA world is one of the most important challenges facing biology today, and the non-coding RNAs within it represent a gold mine of potential new biomarkers and drug targets.
- lncRNAs can be transcribed as whole or partial natural antisense transcripts (NAT) to coding genes, or located between genes or within introns. Some lncRNAs originate from pseudo genes (Milligan & Lipovich, 2015).
- lncRNAs may be classified into different subtypes (Antisense, Intergenic, Overlapping, Intronic, Bidirectional, and Processed) according to the position and direction of transcription in relation to other genes (Peschansky & Wahlestedt, 2014, Mattick & Rinn, 2015).

- lncRNA expression- Gene expression profiling and in situ hybridization studies have revealed that lncRNA expression is developmentally regulated, can be tissue- and cell-type specific, and can vary spatially, temporally, or in response to stimuli.
- Many lncRNAs are expressed in a more tissue-specific fashion and with greater variation between tissues compared to protein-coding genes (Derrien et al., 2012).

3.2.7. Gene ontology (GO)

Gene ontology is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. More specifically, the project aims to:

- Maintain and develop its controlled vocabulary of gene and gene product attributes;
- Annotate genes and gene products, and assimilate and disseminate annotation data; and
- Provide tools for easy access to all aspects of the data provided by the project, and to enable functional interpretation of experimental data using the GO, for example via enrichment analysis.

3.2.8. Differential expression

Based on the embryological evidence for genomic equivalence (and on bacterial models of gene regulation), a consensus emerged in the 1960s that cells differentiate through differential gene expression. The three postulates of differential gene expression are as follows:

- Every cell nucleus contains the complete genome established in the fertilized egg. In molecular terms, the DNAs of all differentiated cells are identical.
- The unused genes in differentiated cells are not destroyed or mutated, and they retain the potential for being expressed.
- Only a small percentage of the genome is expressed in each cell, and a portion of the RNA synthesized in the cell is specific for that cell type.

3.2.9 Tumour expression

- A more powerful result of gene expression profiling is the ability to further classify tumours into subtypes having distinct biological properties and impact on prognoses.

- For example, some diffuse large B-cell lymphomas (DLBCLs) are indistinguishable based on histological methods yet are clinically heterogeneous: 40% of patients respond well and exhibit prolonged survival while the remaining 60% do not.

3.2.10. Mitochondria role in Cancer

- Mitochondria plays important roles in cellular energy metabolism, free radical generation, and apoptosis. Defects in mitochondrial function have long been suspected to contribute to the development and progression of cancer.
- A key event in carcinogenesis involved the development of an "injury" to the respiratory machinery, resulting in compensatory increases in glycolytic ATP production.
- Eventually, malignant cells would satisfy their energy needs by producing a large portion of their ATP through glycolytic mechanisms rather than through oxidative phosphorylation. Due to the inherent inefficiency of glycolytic ATP generation, this represents a somewhat unique metabolic state of the malignant cells and would require high consumption of glucose to fulfill cellular energy requirements.
- This is in contrast to many normal cells, which use oxidative phosphorylation as the preferred means of ATP generation with high efficiency. The differences in energy metabolism between normal and cancer cells constitute a biochemical basis to speculate that therapeutic strategies might be developed to selectively kill cancer cells due to their inherently compromised respiratory state.

CHAPTER4 SOURCES OF DATA

As stated earlier, we have collected our data from previously existing cancer databases in order to create a knowledge base. We are trying to collect all information under one roof to make it user friendly and to save time searching for multiple cancers in separate databases.

Following are some of the databases from which we have collected and gained some very useful data:

DATA	DATA SOURCE	URL
Histology of cancer	COSMIC (Catalogue Of Somatic Mutations In Cancer)	http://cancer.sanger.ac.uk/cosmic/download
Gene ontology	CCGD (Candidate Cancer Gene Database)	http://ccgd-starrlab.oit.umn.edu/download.php
miRNA and Long non-coding rna	microRNA Cancer Association Database	http://mircancer.ecu.edu/download.jsp
cancer drugs	National Cancer Institute (NIH)	www.cancer.gov
Cancer Genomics Data	Omicsoft corporation	http://www.omicsoft.com/
Information related to cancer genes	Network of Cancer Genes	http://nCG.kcl.ac.uk/download.php
MiRNA data	OncomiRDB	http://bioinfo.au.tsinghua.edu.cn/member/jgu/oncomirdb/
SNP	Tumor Portal	http://www.tumorportal.org/

CHAPTER 5 TECHNOLOGY STACK

6.1. Components used in making of the database and the GUI:

6.1.1 HTML and CSS

- HyperText Markup Language, commonly referred to as HTML, is the standard markup language used to create web pages.
- Along with CSS, and JavaScript and PHP, HTML is a cornerstone technology, used by most websites to create visually engaging webpages, user interfaces for web applications.
- Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language.
- CSS is designed primarily to enable the separation of document content from document presentation, including aspects such as the layout, colors, and fonts. This separation can improve content accessibility, provide more flexibility and control

6.1.2 Bootstrap

- **Bootstrap** is a free and open-source front-end library for creating websites and web applications. It contains HTML and CSS-based design templates for typography, forms, buttons, navigation and other interface components, as well as optional JavaScript extensions. It aims to ease the development of dynamic websites and web applications.
- Bootstrap is a front end web framework, that is, an interface for the user, unlike the server-side code which resides on the "back end" or server.
- Bootstrap is the second most-starred project on GitHub, with over 95 thousand stars and more than 40 thousand forks.

6.1.3 MySQL 5.6.24

- It is an open-source relational database management system (RDBMS).
- It is the world's second most widely used RDBMS, and the most widely used open-source client–server model RDBMS.
- MySQL is a popular choice of database for use in web applications, and is a central component of the widely used LAMP open source web application software stack.

6.1.4 PHP 5.6

- PHP is a server-side scripting language designed for web development but also used as a general-purpose programming language.
- PHP code may be embedded into HTML code, or it can be used in combination with various Web template systems and web frameworks.
- PHP code is usually processed by a PHP interpreter implemented as a module in the web server or as a Common Gateway Interface (CGI) executable. The web server combines the results of the interpreted and executed PHP code, which may be any type of data, including images, with the generated web page.

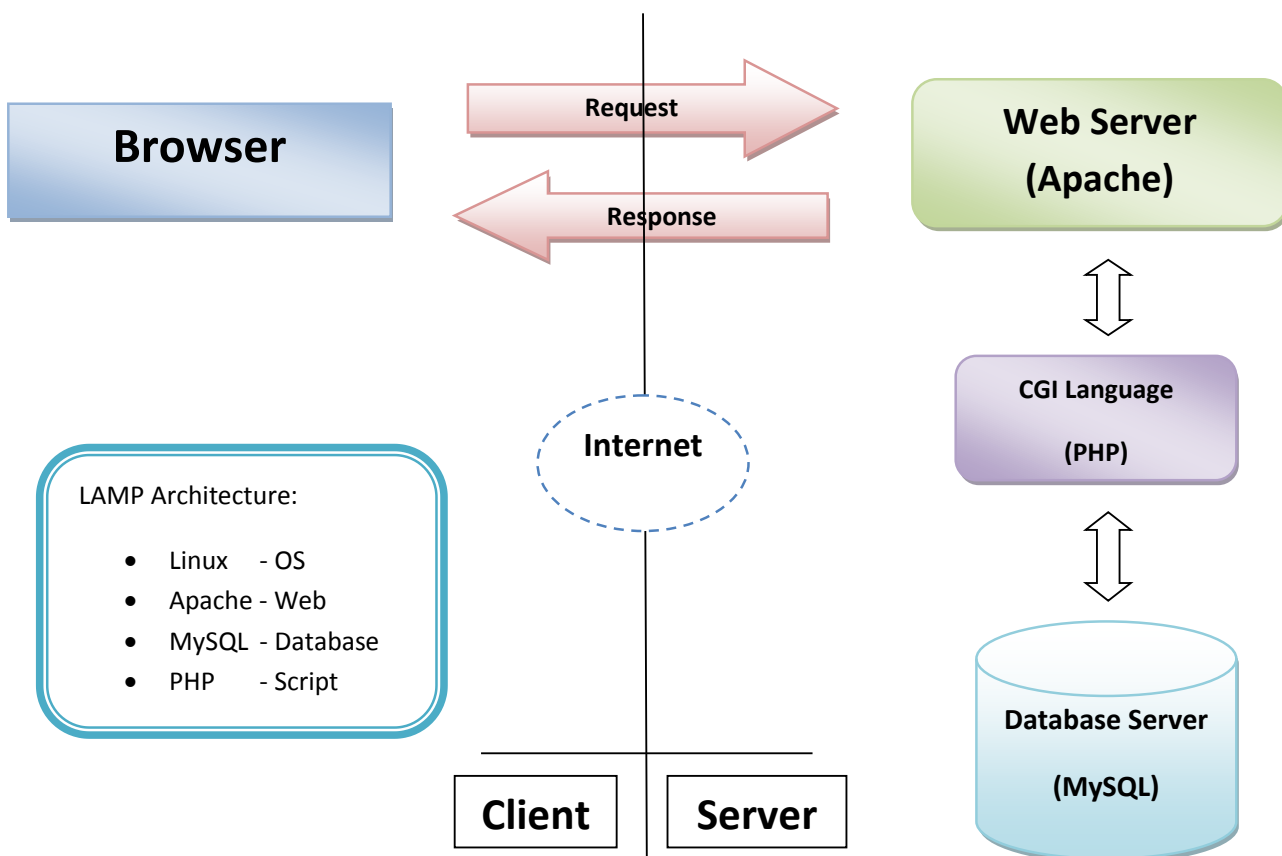
6.2. System architecture and requirements

6.2.1 Lamp Stack

- LAMP is an archetypal model of web service solution stacks, named as an acronym of the names of its original four open-source components: the Linux operating system, the Apache HTTP Server, the MySQL relational database management system (RDBMS), and the PHP programming language.
- The LAMP components are largely interchangeable and not limited to the original selection. As a solution stack, LAMP is suitable for building dynamic web sites and web applications.
- Since its creation, the LAMP model has been adapted to other component, though typically consisting of free and open-source software. For example, an equivalent installation on the Microsoft Windows family of operating systems is known as WAMP.

There is a client side as well as a server side. Client side is used by the user through the browser e.g. Firefox. And the request of the user is sent to the server side which again displays the information using table format which is extracted by the database (MySQL) with the help of our main CGI language i.e. PHP.

Fig 7 - Overall backend process can be explained by the following diagram



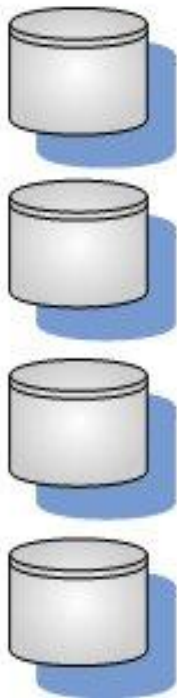
6.2.2 OLTP vs. OLAP

We can divide IT systems into transactional (OLTP) and analytical (OLAP). In general we can assume that OLTP systems provide source data to data warehouses, whereas OLAP systems help to analyze it.

- **OLTP (On-line Transaction Processing)** is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE). The main emphasis for OLTP systems is put on very fast query processing, maintaining data integrity in multi-access environments and an effectiveness measured by number of transactions per second. In OLTP database there is detailed and current data, and schema used to store transactional databases is the entity model (usually 3NF).
- **OLAP (On-line Analytical Processing)** is characterized by relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems a response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. In OLAP database there is aggregated, historical data, stored in multi-dimensional schemas (usually star schema).

Entity Relational Models

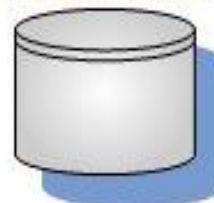
OLTP Systems



Dimensional Models

OLAP Systems

Enterprise Data Warehouses



Data Marts

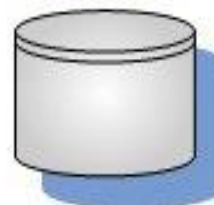


Fig 8: OLTP and OLAP (Source- Dale Anderson - Big Data, Database Migration)

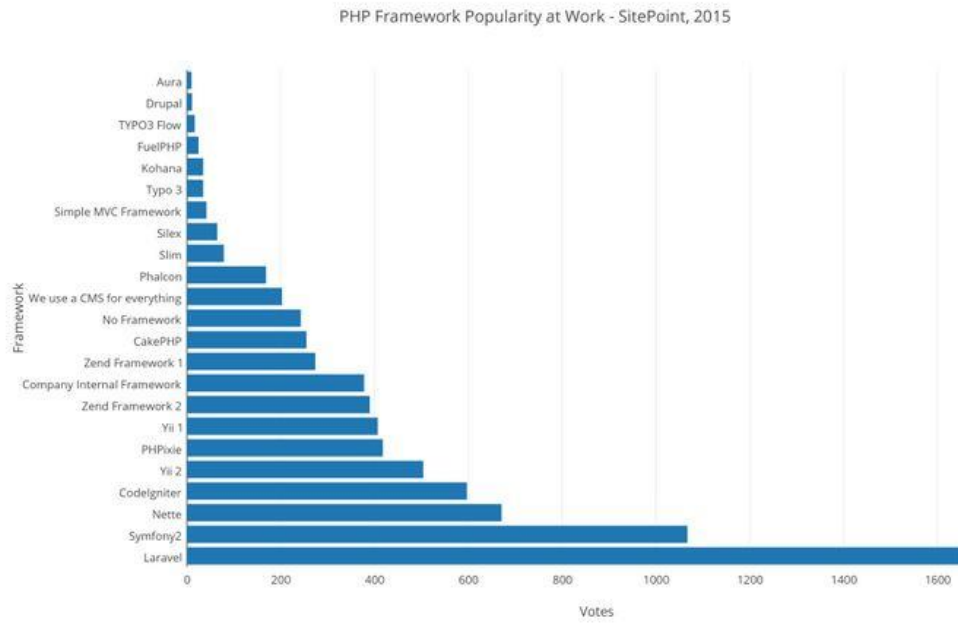


Fig 9: PHP framework popularity

Data Source : Bruno Skvorc-The Best PHP Framework for 2015: SitePoint Survey Results

CHAPTER 6

WEB DESIGN

We have weaved the front end using some of the latest end to end technologies.

- HTML5
- CSS3
- JAVASCRIPT

We've worked with the latest twitter's frond end framework Bootstrap (link)

Design Principles:

Our Design pattern includes

- Z pattern design



Fig 10- Z pattern

- Call to action button
- Homepage slider

The design sense has been adopted from the Google's material design.

The name of our database is **CANCERTOME**.

7.1 Front End

This is the **main page** of our GUI.

- We have provided links for all of our cancer types with each page separately defining the cancers.
- On the top left corner we have a search bar which will allow the users to search for any type of information they want of the desired cancer.
- We have a navigation bar which is connected to additional pages of :
 - Download
 - Contact
 - Home
 - About



For information about how to effectively use this database, have a look at the about

General Information

Gene ID:

Cancer type:

Gene Name:

Network Analysis

miRNA from literatures

lncRNA

Name of lncRNA:

Cancer type:

Gene ID of lncRNA:

Type of activity:

Morphological data

p53 mutations

Fig 11: Home Page of GUI

- Search options provided in our knowledgebase are as follows:

General Information

Gene ID:

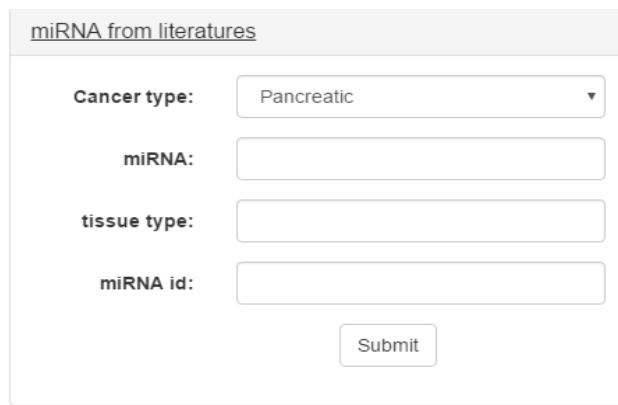
Cancer type:

Gene Name:

Network Analysis

Cancer type:

Gene Name:



The image shows a search form titled "miRNA from literatures". It contains four input fields: "Cancer type:" with a dropdown menu showing "Pancreatic", "miRNA:", "tissue type:", and "miRNA id:". A "Submit" button is located at the bottom center of the form.

Fig 12 – Search options

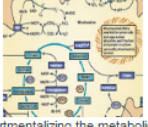
- Craftsmen use of JavaScript can be seen as the the bootstrap’s Accordion has been implemented which provides the user with an excellent User Experience (UX) on desktop as well as on mobile devices.
- Search options work independently as well as collectively.
For eg - In the table ‘General Information’ you can solely enter the Gene ID and search by this parameter or you can enter the Gene ID with the Cancer Type and Gene Name to make the search more specific.
- More information is provided on the main page separately for :
 - Mitochondrial Data
 - Tumor expression Data
- To narrow down the search related to mitochondrial information, search has been provided separately for it.
- The results are generated in a tabular format.
- Sample table is given below. For more information visit the website www.cancertome.online

Sr no.	Gene Id	Cancer type	Number of studies	Pubmed ID	Effect of mutation	Gene
1	5728	Pancreatic Cancer	34	22421440	Loss	PTEN
2	5728	Pancreatic Cancer	34	22699621	Not Determined	PTEN
3	5728	Pancreatic Cancer	34	22699621	Not Determined	PTEN
4	1387	Pancreatic Cancer	25	22421440	Loss	CREBBP
5	1387	Pancreatic Cancer	25	22699621	Not Determined	CREBBP
6	1387	Pancreatic Cancer	25	22699621	Not Determined	CREBBP
7	57178	Pancreatic Cancer	21	22699621	Not Determined	ZMIZ1
8	57178	Pancreatic Cancer	21	22699621	Not Determined	ZMIZ1
9	7403	Pancreatic Cancer	20	22421440	Loss	KDM6A

Fig 13 – Sample result page

Cancerome Home About [Download](#) [Contact](#) ▾

Mitochondrial information related to cancer



Mitochondria undertake multiple critical functions in a cell. In addition to compartmentalizing the metabolic pathways and physiological states of the cell, the mitochondria generate much of the cellular energy, regulate the cellular redox state, produce most of the cellular reactive oxygen species (ROS), buffer cellular Ca²⁺ and initiate cellular apoptosis. Mitochondria were first proposed to be relevant to cancer by Otto Warburg who reported that cancer cells exhibited "aerobic-glycolysis". Although this was originally interpreted as indicating that the function of mitochondria was defective, we now understand that cancer cells are in an altered metabolic state. Little attention has been paid to the potential mutations that can affect mitochondrial function in cancer, outside of specific mutations in genes such as fumarate hydratase (FH), which is associated with a rare cancer syndrome, hereditary leiomyomatosis and renal cell carcinoma (HLRCC). Genes such as FH are encoded in the nuclear DNA, so once mutated they affect all mitochondria in a cell. However, mitochondria also have their own DNA and mutations in mitochondrial genes are common and have been shown to be involved in human diseases, such as mitochondrial myopathy. The resurgence of interest in metabolism in cancer cells has started to focus attention back on the mitochondria and there is increasing evidence that mutations in mitochondrial DNA encoded genes can contribute to the development of cancer. It is possible that such mutations provide metabolic adaptivity to the cancer cell. This poster, covers our current understanding of the contribution of mitochondrial function to cancer cell metabolism. The editorial content of this poster was conceived and developed by the poster author Douglas C. Wallace, and the editorial team at Nature Reviews Cancer.

Search Form

Cancer type:

Mutated Gene / Region:

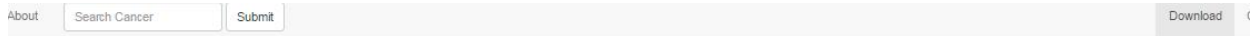
Fig 14 – Mitochondrial search page

We have designed a separate page for Mitochondria alterations in Cancer. There is a form where you can enter the cancer type and mutated gene or region like D-loop, cyt b etc. The result is in a table format.

There is another button with the title “More Information” with which you can open a table with mitochondrial alterations in reference to their Pubmed Ids.

To help our users further our data is downloadable and download links have been provided.

The data can be downloaded in CSV format.



This utility helps you to download the data as mysql dump file (.csv format)

1. Complete Database dump
2. General Information about cancers
3. Tumor Expression Data
4. p53 mutations data
5. lncRNA data
6. mitochondrial data

Fig 15 – Download page

Following is the Contact Us page

Hello, Please fill in this form so that we can contact you as soon as possible.

Name:

Email:

Mobile:

Fig 16 – Contact page

7.2 Back End

A "back-end" application or program serves indirectly in support of the front-end services, usually by being closer to the required resource or having the capability to communicate with the required resource. The back-end application may interact directly with the front-end or, perhaps more typically, is a program called from an intermediate program that mediates front-end and back-end activities.

In backend processing we have used :

- MySQL
- PHP
- PHP's Codeigniter MVC framework

Model-view-controller (MVC) is a software architectural pattern for implementing user interfaces on computers. It divides a given software application into three interconnected parts, so as to separate internal representations of information from the ways that information is presented to or accepted from the user.

Traditionally used for desktop graphical user interfaces (GUIs), this architecture has become extremely popular for designing web applications.

We can define, structure and apply MySQL queries to our tables using phpMyAdmin/MySQL. phpMyAdmin is a free and open source tool written in PHP intended to handle the administration of MySQL or MariaDB with the use of a web browser. It can perform various tasks such as creating, modifying or deleting databases, tables, fields or rows; executing SQL statements; or managing users and permissions.

- Sample structure of our database containing 11 tables according to the attributes considered.
You can explore any table by clicking on the desired table name.

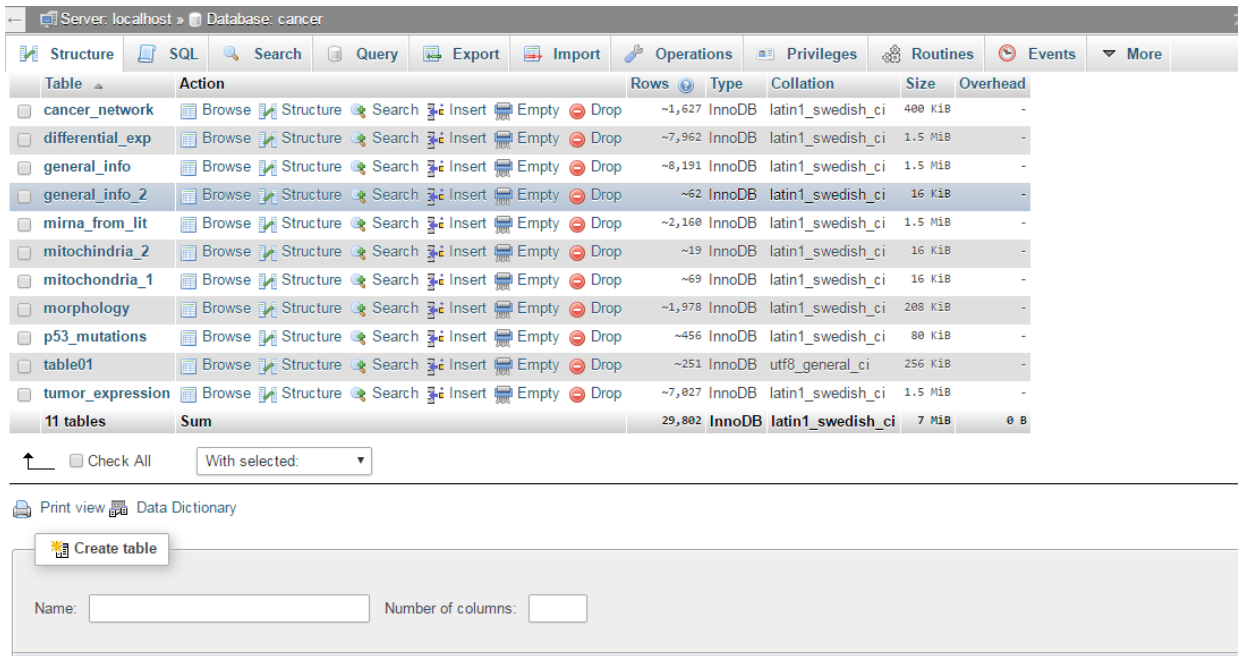


Fig 17 – phpMyadmin Database

- The structure of our 1st table “cancer_network”. You can see all the attribute names with their type and many more options associated to it.

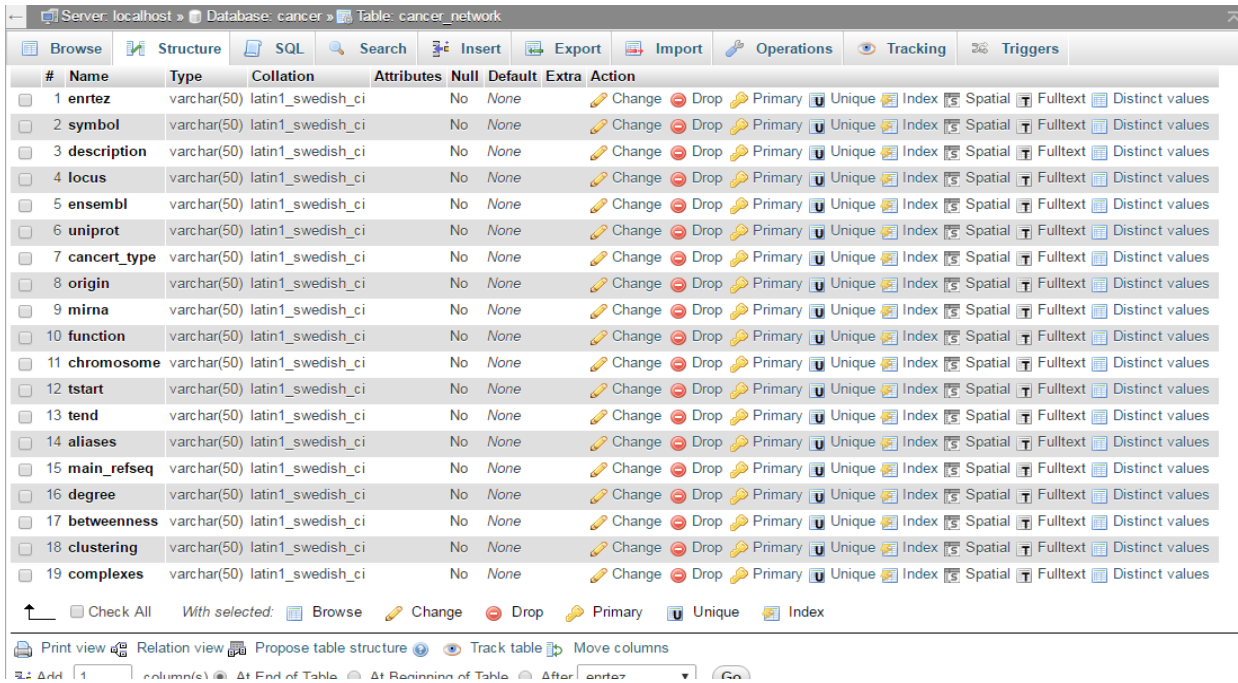


Fig 18 – Sample table structure

- On browsing the above table “cancer_network” you can see the various entries.
- MySQL queries can be applied to search accordingly.

enrtz	symbol	description	locus	ensembl	uniprot	cancer_type	origin	mirna	function
25	ABL1	ABL proto-oncogene 1, non-receptor tyrosine kinase	9q34.1	ENSP00000361423	P00519	cgc	Eukaryotes	3	Cell cycle, Cellular metabolism, Cellular processes,
27	ABL2	ABL proto-oncogene 2, non-receptor tyrosine kinase	1q25.2	ENSP00000427562	P42684	cgc	Eukaryotes	5	Cellular metabolism, Cellular processes, Regulation
60	ACTB	actin, beta	7p22	ENSP00000224784;ENSP00000295137;ENSP00000326022;EN	P60709	can	Last Universal Common Ancestor	2	Cell response to stimuli
90	ACVR1	activin A receptor, type I	2q23-q24	ENSP00000263640	Q04771	can	Metazoans	8	Cell cycle, Cell motility and interactions, Cell res
91	ACVR1B	activin A receptor, type IB	12q13	ENSP00000442656	P36896	can	Metazoans	6	Cell cycle, Cellular metabolism, Cellular processes,
92	ACVR2A	activin A receptor, type IIA	2q22.3	ENSP00000241416	P27037	can	Metazoans	5	Cellular metabolism, Cellular processes, Development
102	ADAM10	ADAM metallopeptidase domain 10	15q22	ENSP00000260408	O14672	can	Opisthokonts	2	Cell motility and interactions, Cellular metabolism
107	ADCY1	adenylate cyclase 1 (brain)	7p13-p12	ENSP00000297323	Q08828	can	Last Universal Common Ancestor	0	Cellular metabolism, DNA/RN metabolism and transcr
114	ADCY8	adenylate cyclase 8 (brain)	8q24	ENSP00000260600;ENSP00000286355	P40145	can	Last Universal Common Ancestor	0	Cellular metabolism, DNA/RN metabolism and transcr
125	ADH1B	alcohol dehydrogenase 1B (class I)	4q23	ENSP00000306606	P00325	can	Last Universal Common Ancestor	0	Cellular metabolism

Fig 19 – Sample table data through phpMyAdmin

Chapter 7

CHALLENGES

- Data increases at an exponential rate therefore regular updating of data is required.
- The main challenge that we face is the ever increasing data.
- Regular updating of the front end is also important.
- Maintenance of the the backend and the overall process is also needed from time to time.
- We cannot rely on our application to enforce all data integrity rules. A SQL Server database is designed to enforce some data integrity rules.
- Using external resources for data has to be formatted so that it fits our requirement of our database.
- We do not want our database to run out of disk space. To keep your disk drives from filling up we need to monitor them to ensure there is adequate disk space on our physical drives. Ideally, we should track the amount of disk space growth over time.
- Maintaining a stable production environment is extremely important. Database changes need to be well thought out and planned. Having a change management process provides some structure around how changes are made. We need to develop a change management process for our environment.
- It is better to encapsulate our application code into stored procedures. This will allow us to only give users EXECUTE permissions. Doing this means users will not be able to access our tables directly outside the applications using SELECT, INSERT, UPDATE, and DELETE statements.

WEB RESOURCES

- <http://ccgd-starrlab.oit.umn.edu/>
- <http://cancer.sanger.ac.uk/cosmic/download>
- <http://mircancer.ecu.edu/downloads/>
- www.cancer.gov
- www.tumorportal.org/
- http://interactome.dfci.harvard.edu/H_sapiens/index.php
- <https://www.digitalocean.com/community/tutorials/5-common-server-setups-for-your-web-application>
- <http://datawarehouse4u.info/OLTP-vs-OLAP.html>
- <http://www.sitepoint.com/best-php-framework-2015-sitepoint-survey-results/>
- <https://google.com/images>
- <http://ncg.kcl.ac.uk/>
- <http://bioinfo.au.tsinghua.edu.cn/member/jgu/oncomirdb/>
- <https://hive.biochemistry.gwu.edu/tools/bioexpress/>
- <http://www.ocimumbio.com/bioexpress-system/>
- <http://www.ncri.org.uk/what-we-do/research-database/>
- <https://tcga-data.nci.nih.gov/tcga/>

REFERENCE ARTICLES

[1] Pavlopoulou A, Spandidos DA, Michalopoulos I, "Human cancer databases (review)", Center of Systems Biology, Academy of Athens(Athens), Laboratory of Clinical Virology, Medical School, University of Crete, doi: 10.3892/or.2014.3579

[2] Petitjean A, Mathe E, Kato S, et al. Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Hum Mutat.*2007;28:622–629. doi: 10.1002/humu.20495

[3]Carew JS, Huang P, "Mitochondrial defects in cancer", Department of Molecular Pathology, The University of Texas, The Graduate School of Biomedical Sciences, University of Texas Health Sciences Center, doi: 10.1186/1476-4598-1-9

[4]Tan DJ, Bai RK, Wong LJ. Comprehensive scanning of somatic mitochondrial DNA mutations in breast cancer. *Cancer Res.* 2002;62:972–976.

[5]Sharp MG, Adams SM, Walker RA, Brammar WJ, Varley JM. Differential expression of the mitochondrial gene cytochrome oxidase II in benign and malignant breast tissue. *J Pathol.* 1992;168:163–168

[6]Savre-Train I, Piatyszek MA, Shay JW. Transcription of deleted mitochondrial DNA in human colon adenocarcinoma cells. *Hum Mol Genet.* 1992;1:203–204

[7]Maximo V, Soares P, Seruca R, Rocha AS, Castro P, Sobrinho-Simoes M. Microsatellite instability, mitochondrial DNA large deletions, and mitochondrial DNA mutations in gastric carcinoma. *Genes Chromosomes Cancer.* 2001;32:136–143. doi: 10.1002/gcc.1175

[8]Parrella P, Xiao Y, Fliss M, Sanchez-Cespedes M, Mazzarelli P, Rinaldi M, Nicol T, Gabrielson E, Cuomo C, Cohen D, Pandit S, Spencer M, C Rabitti, Fazio VM, Sidransky D. Detection of mitochondrial DNA mutations in primary breast cancer and fine-needle aspirates. *Cancer Res.*2001;61:7623–7626.

[9]Polyak K, Li Y, Zhu H, Lengauer C, Willson JK, Markowitz SD, Trush MA, Kinzler KW, Vogelstein B. Somatic mutations of the mitochondrial genome in human colorectal tumours. *Nat Genet.*1998;20:291–293. doi: 10.1038/3108

[10]Jones JB, Song JJ, Hempen PM, Parmigiani G, Hruban RH, Kern SE. Detection of mitochondrial DNA mutations in pancreatic cancer offers a "mass"-ive advantage over detection of nuclear DNA mutations. *Cancer Res.* 2001;61:1299–1304