Nijhum Paul

Abstract

There are many resources available that are dedicated to different aspects of cancer. However, this information is scattered to different publicly available domain. With the intention of getting multiple aspects of different cancers easily, I have created a database that contains different facts of cancers, like genes related to a cancer, recent AI tools available for each cancer type, survival rate, and so on. I have used multiple tools like python, PostgreSQL, SQL, etc. to build this project. This database can answer many questions related to cancers, like which cancer has the most/ least survival rate, which genes are responsible for a particular cancer type, which cancer type has the most recently created AI tool, etc. This database can be helpful for general users and researchers by providing all these information in a short time.

Introduction

Integrated data of a particular area is a crucial tool in research as it provides knowledge of a specific area that helps general people/ researchers using them for further analysis. There has been numerous research, case study, treatment, new technology, and invention in the field of cancer. Research in cancer is transforming and saving many lives for a long time. The goal of studying cancer is to develop safe and effective methods to prevent, detect, diagnose, treat, and cure cancer.

There have been several projects on integrating the cancer information and make them available online by researchers. Thomas et. al. described the development of a web-based resource, Pancreatic Cancer Database (PCD) as a unified platform for pancreatic cancer research. It contains manually curated information pertaining to quantitative alterations in miRNA, mRNA, and proteins obtained from small-scale as well as high-throughput studies of pancreatic cancer tissues and cell lines. Sarver et. al. developed the OncomiR Cancer Database (OMCD), hosted on a web server, which allows easy and systematic comparative genomic analyses of miRNA sequencing data derived from more than 9500 cancer patients tissue samples available in the Cancer Genome Atlas (TCGA). OMCD includes associated clinical information and is searchable by organ-specific terms common to the TCGA.

In my project, I have created a database that integrates different aspects of multiple cancer types. This database is a reliable and easy-to-use repository to gain information like relevant genes, treatment, survival rate, and existing AI tools for a particular cancer type (figure 1).

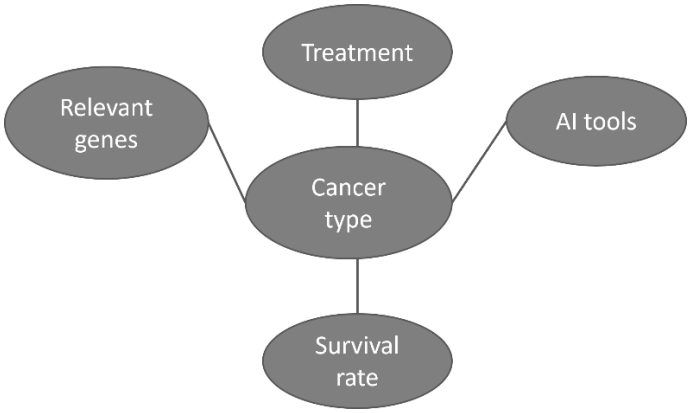

Figure 1: Functionality of integrated cancer database.

I have collected gene data from COSMIC [3], TSGene database [4], and two cancer related articles [5][6]. These data collection was performed by python. I also collected other data like, gene description and the resources I got the information from. I also collected information of different cancer types, their survival rate, and treatment. Different cancer prediction tools, their creation date and author name were also collected. These data collection was manually curated. Then I used sql to create a database and enter all data into it. I loaded the tables with data in PostgreSQL database using psycopg2. Then I ran sone queries that answers multiple questions regarding to cancers, genes and AI tools.

Course relevance

Entity-Relationship Modeling

To translate the real-world expectation into a data model, I have created an entity-relationship modeling based on my project. It is the basic design upon which my database will be built. I have used three entities, which are genes, cancers and AI tools. Each of these entities have their own attributes.
Genes have attributes like ID, name, resources, etc. The primary key of genes is gene_ID, which is unique for each gene.
Cancer entity has attribute like ID, name, treatment, survival rate, etc. cancer_ID is primary key and is unique for each cancer type.
aiTools has ID, name, creation date and created by attributes. tool_id is the primary key and it is unique for each tool.

ER-to-Relational Mapping

The ER-model is translated to a relational model. I have explained the primary and foreign keys of each entity here. I also indicated the cardinality ratio of each relation to show if a relation is one-to-many, many-to-one, one-to-one, or many-to-many relation. Genes and cancer have partial participation among them, so I have mapped them as many-to-many relationship. One gene can relate to many cancer types and one cancer type can relate to many cancers. Also, one gene may not relate to any cancer type, but each cancer type should be related to some genes. Cancer and aiTools have one-to-many relationship, because multiple tools can be created for just one cancer type. Also, it is possible that one cancer type does not have any aiTool, but each aiTools must be related to a cancer type.

SQL

Structured Query Language (SQL) is used to manage relational databases and perform various operations on the data in them. I have used SQL for table creation and data manipulation. I have used query CREATE TABLE table_name to create table. In each table, all attributes along with their data type is specified. I also indicated if a key is primary or foreign key. The foreign key is mentioned along with the table it references.
For data entry, I used INSERT INTO table_name query.
I also ran other queries like joining tables, sorting tables, and so on. Some most commonly used queries like SELECT, JOIN, ORDER BY, COUNT were used.

Implementation

Entity- relation diagram (figure 2) shows three entities, which are genes, cancers and aiTool. Genes and cancer have partial participation among them, so I have mapped them as many-to-many relationship. One gene can relate to many cancer types and one cancer type can relate to many cancers. For many-to-many relationship, a separate table gene_cancer had to be created. Also, one gene may not relate to any cancer type, but each cancer type should be related to some genes. Cancer and aiTools have one-to-many relationship, because multiple tools can be created for just one cancer type.

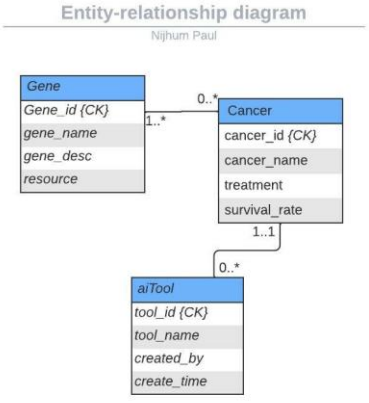

Figure 2: ER diagram

Figure 3 shows the relational diagram that contains three entities, which are genes, cancers and aiTool. Genes have attributes like ID, name, description and resource. The primary key of genes is gene_ID, which is unique for each gene. Cancer entity has attribute like ID, name, treatment, and survival rate. cancer_id is primary key and is unique for each cancer type. aiTools has ID, name, creation date and created by attributes. tool_id is the primary key and it is unique for each tool.

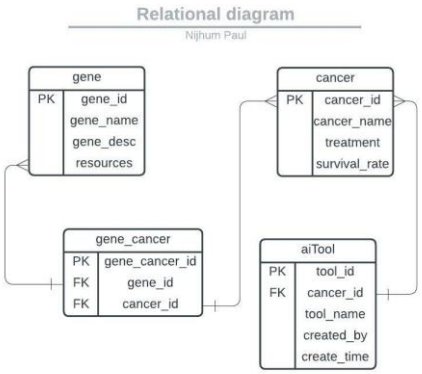

Figure 3: Relational diagram

SQL query is used (figure 4) for table creation and data manipulation. I have used query CREATE TABLE table_name to create table. In each table, all attributes along with their data type is specified. I also indicated if a key is primary or foreign key. The foreign key is mentioned along with the table it references. For data entry, I used INSERT INTO table_name query.

```
assign1.sql - NIJH...(NIJHUM\nijhu (53))   projsql.sql - NIJHU...(NIJHUM\nijhu (51))*

CREATE TABLE genes(
gene_id INT NOT NULL,
gene_name VARCHAR(200),
gene_desc VARCHAR(500),
resources VARCHAR(200),
PRIMARY KEY(gene_id)
);

INSERT INTO genes(gene_id,gene_name,gene_desc, resources)
VALUES
(1,'MIR375','microRNAs (miRNAs) are short (20-24 nt) non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multic
(2,'ARMC10','This gene encodes a protein that contains an armadillo repeat and transmembrane domain. ','Genecards.org'),
(3,'ANXA1','This gene encodes a membrane-localized protein that binds phospholipids.','Genecards.org'),
(4,'MIR504','microRNAs (miRNAs) are short (20-24 nt) non-coding RNAs that are involved in post-transcriptional regulation of gene expression in multic
(5,'CD79A','The B lymphocyte antigen receptor is a multimeric complex that includes the antigen-specific component, surface immunoglobulin (Ig).','Gen
(6,'HTRA2','This gene encodes a serine protease. The protein has been localized in the endoplasmic reticulum and interacts with an alternatively splic
(7,'LLGL1','This gene encodes a protein that is similar to a tumor suppressor in Drosophila. The protein is part of a cytoskeletal network and is asso
(8,'JAK1','This gene encodes a membrane protein that is a member of a class of protein-tyrosine kinases (PTK) characterized by the presence of a secon

CREATE TABLE cancer(
cancer_id INT NOT NULL,
cancer_name VARCHAR(200),
treatment VARCHAR(200),
survival_rate VARCHAR(200),
PRIMARY KEY(cancer_id)
);


INSERT INTO cancer(cancer_id, cancer_name, treatment, survival_rate)
VALUES
(101, 'breast cancer', 'Surgery, Radiation, Chemotherapy, Hormone therapy', '90%'),
(102, 'pancreatic cancer', 'Surgery, Radiation,Ablation and Embolization, Chemotherapy', '11%'),
(103, 'skin cancer', 'Radiation, Chemotherapy, Freezing, Excisional surgery', '93%'),
(104, 'lung cancer', 'surgery, chemotherapy, radiation therapy, targeted therapy', '18.6%'),
(105, 'prostate cancer', 'Surgery, Radiation, Chemotherapy, Immunotherapy', '98%'),
(106, 'kidney cancer', 'Surgery, Radiation, Chemotherapy, Ablation and Other Local Therapy', '76%'),
(107, 'colorectal cancer', 'Surgery, Radiation, Chemotherapy, Immunotherapy', '65%');

CREATE TABLE gene_cancer(
```

Figure 4: SQL query

Result

I have run some queries like joining two tables, sorting cancer types based on their survival rate, sorting AI tools based on their creation time, and so on. Some most commonly used queries like SELECT, JOIN, ORDER BY, COUNT was used.

I have joined two tables, genes and cancer (figure 5). This new table shows which cancer is related to which gene and vice versa. The description of a gene also indicates which properties of a gene can be responsible for a cancer type. For example, JAK1, CD79A and HTRA2 genes are responsible for prostate cancer.

	gene_name	gene_desc	cancer_name
1	MIR375	microRNAs (miRNAs) are short (20-24 nt) non-codin...	breast cancer
2	ARMC10	This gene encodes a protein that contains an armadi...	colorectal cancer
3	ANXA1	This gene encodes a membrane-localized protein th...	skin cancer
4	MIR504	microRNAs (miRNAs) are short (20-24 nt) non-codin...	lung cancer
5	CD79A	The B lymphocyte antigen receptor is a multimeric co...	prostate cancer
6	CD79A	The B lymphocyte antigen receptor is a multimeric co...	pancreatic cancer
7	JAK1	This gene encodes a membrane protein that is a me...	kidney cancer
8	JAK1	This gene encodes a membrane protein that is a me...	prostate cancer
9	HTRA2	This gene encodes a serine protease. The protein ha...	prostate cancer

Figure 5: Joining genes and cancer tables.

Figure 6 shows which genes are related to maximum/ minimum number of cancer types. It shows that JAK1 and CD79A are related to two different cancer types, while others relate to only one cancer type.

	gene_id	gene_name	number of cancers per gene
1	1	MIR375	1
2	2	ARMC10	1
3	3	ANXA1	1
4	4	MIR504	1
5	6	HTRA2	1
6	8	JAK1	2
7	5	CD79A	2

Figure 6: Number of cancer types related to a gene.

I have sorted the aiTool table based on creation time (figure 7). This tables shows that the tool for lung cancer and colon cancer are the most recently published AI tools that were published in 2022. They were created by Memorial Sloan Kettering Cancer Center and The University of Texas MD Anderson Cancer Center, respectively.

	tool_id	cancer_id	tool_name	created_by	create_time
1	1001	101	PREDICT	Ewan Gray	2018
2	1003	103	cSCCscore	WeiWang	2018
3	1002	102	IRE	Chaobin He	2020
4	1005	105	PREDICT Prostate	Gaëtan Devos	2020
5	1006	106	Colon Cancer Survival Calculator	The University of Texas MD Anderson Cancer Center	2022
6	1004	104	Lung Cancer Risk Assessment Tool	Memorial Sloan Kettering Cancer Center	2022

Figure 7: Sorted aiTool table based on creation time.

Figure 8 shows the sorted cancer table based on survival rate of the cancer types. It indicates that prostate cancer has the most survival rate and pancreatic cancer has the least survival rate. Breast (90%) and skin cancers (93%) have almost the

same survival rate. Pancreatic cancer has the least survival rate. This table also shows that most of the cancer has some common treatment types, which are surgery, radiation, and chemotherapy.

	cancer_id	cancer_name	treatment	survival_rate
1	102	pancreatic cancer	Surgery, Radiation,Ablation and Embolization, Che...	11%
2	104	lung cancer	surgery, chemotherapy, radiation therapy, targeted t...	18.6%
3	107	colorectal cancer	Surgery, Radiation, Chemotherapy, Immunotherapy	65%
4	106	kidney cancer	Surgery, Radiation, Chemotherapy, Ablation and Ot...	76%
5	101	breast cancer	Surgery, Radiation, Chemotherapy, Hormone therapy	90%
6	103	skin cancer	Radiation, Chemotherapy, Freezing, Excisional sur...	93%
7	105	prostate cancer	Surgery, Radiation, Chemotherapy, Immunotherapy	98%

Figure 8: Sorted cancer table based on survival rate.

Conclusion

This database was created in an intention of integrating multiple aspects of different types of cancers. Different tools were used like BeautifulSoup, pandas, PostgreSQL, SQL, psycopg2, etc. Different important information related to cancers can be retrieved using this database, like which cancer has the most/ least survival rate, which genes are responsible for a particular cancer type, which cancer type has the most recently created tool, and so on. However, due to short availability of time, enough data could not be collected. In future, this database can be refined with more features and data, and it can be turned into a web based online tool also.

References

1. Joji Kurian Thomas, Min-Sik Kim, Lavanya Balakrishnan, Vishalakshi Nanjappa, Rajesh Raju, Arivusudar Marimuthu, Aneesha Radhakrishnan, Babylakshmi Muthusamy, Aafaque Ahmad Khan, Sruthi Sakamuri, Shantal Gupta Tankala, Mukul Singal, Bipin Nair, Ravi Sirdeshmukh, Aditi Chatterjee, T S Keshava Prasad, Anirban Maitra, Harsha Gowda, Ralph H Hruban & Akhilesh Pandey (2014) Pancreatic Cancer Database, Cancer Biology & Therapy, 15:8, 963-967, DOI: 10.4161/cbt.29188.
2. Sarver, A., Sarver, A., Yuan, C. et al. OMCD: OncomiR Cancer Database. BMC Cancer 18, 1223 (2018). https://doi.org/10.1186/s12885-018-5085-z.
3. Min Zhao, Jingchun Sun, Zhongming Zhao. TSGene: a web resource for tumor suppressor genes. Nucleic Acids Research, 2013 Jan;41(Database issue):D970-6.
4. cancer.sanger.ac.uk.
5. breastcancer.org.
6. Shiovitz, S, and L A Korde. “Genetics of breast cancer: a topic in evolution.” Annals of oncology : official journal of the European Society for Medical Oncology vol. 26,7 (2015): 1291-9. doi:10.1093/annonc/mdv022.