KEY COMMON GENES OF COLORECTAL CANCER CHARACTERISTICS BASED ON PROTEIN-PROTEIN INTERACTIONS: A BIOINFORMATICS APPROACH

By

NANDINI BHADRA
143-35-837
&
NUHA TASMIAH
143-35-753

A thesis submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Software Engineering

Department of Software Engineering
DAFFODIL INTERNATIONAL UNIVERSITY

Spring – 2019

All rights reserved by Daffodil International University

APPROVAL

This **Thesis** titled "**Key Common Genes of Colorectal Cancer Characteristics Based on Protein-Protein Interactions: A Bioinformatics Approach**", submitted by **Nandini Bhadra**, **ID: 143-35-837 and Nuha Tasmiah**, **ID: 143-35-753** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Software Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan
Professor and Head **Chairman**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Md. Maruf Hassan
Assistant Professor **Internal Examiner 1**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Asif Khan Shakir
Lecturer **Internal Examiner 2**
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Dr. Md. Nasim Akhtar
Professor **External Examiner**
Department of Computer Science and Engineering
Faculty of Electrical and Electronic Engineering
Dhaka University of Engineering & Technology, Gazipur

DECLARATION

We hereby declare that we have taken this thesis under the supervision of **Md. Habibur Rahman, Lecturer, Department of Software Engineering, Daffodil International University.** We also declare that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.

Nandini Bhadra
ID: 143-35-837
Batch : 15th
Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

Nuha Tasmiah
ID: 143-35-753
Batch : 15th
Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

Certified by:

Md. Habibur Rahman

Lecturer

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

ACKNOWLEDGEMENT

First we express our earnest thanks and gratefulness to almighty Allah for His heavenly blessing, which make us possible to complete this thesis successfully.

It is an fortunate opportunity for us as a student of the Department of Software Engineering, one of the exalted academic centers of the Science and Information Technology Faculty of the Daffodil International University, to express us deep feelings of gratitude to the department and to our honorable teachers and also to the department staff.

We are utmost indebted to our honorable supervisor, **Md. Habibur Rahman, Lecturer, Department of Software Engineering, FSIT, Daffodil International University, Dhaka**, for his excellent guidance, inspiration, encouragement and also for through review of our thesis paper. It was not possible for us to complete our thesis paper successfully without his help. We really thanked and gratefully remember those persons who always helped and encouraged us.

Last of all, we would like to thank to our parents who have given us tremendous inspiration and supports. Without their mental and financial supports we would not be able to complete our thesis.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURE

ABSTRACT

Background: Bioinformatics handles living organism records and inspects the information using computer science facilities. As a result of improving bioinformatics tools and resources, it is now possible to design the protein-protein interaction (PPI) network from disease-associated common genes. PPI networks are varied based on the interaction scores, p-values and coefficiency. Thus, this paper discovers genetic association among colorectal cancer (CRC) characteristics and identifies the statistically highest key common genes of CRC characteristics based on the PPI networks.

Objective: More than a generation humans are being killed by means of CRC characteristics globally. CRC is normally recognized as bowel most cancers or rectal most cancers. CRC characteristics are the main reason for deaths globally. Characteristics of CRC are frequently occurring across the earth and growing due to the fact of frequent threat elements gradually. Common risk elements illnesses have genetic association circuitously or directly. A disease is an abnormal condition in a single gene that affects physique negatively. Biomolecule or protein is the best key for disease renovation. Data mining as properly as data analysis is essential in bioinformatics to locate the favored data.

Results: Knowledge discovery in databases (KDD) process is applied to find out the 6 common disease-associated genes from gene dataset using R. In particular, STRING tool used to investigate the PPI networks and identified highly 3 significant common genes of CRC characteristics.

Conclusions: This study claimed to find out highly significant common genes for CRC characteristics. The genes associated with CRC characteristics are collected from NCBI database using R. To obtain the aim String is used as a tool.

Keywords: PPI, Colorectal Cancer, String, Common gene Identification, R, Data mining

CHAPTER 1

INTRODUCTION

1.1 Background

Bioinformatics is a bunk within the world of science for the present. Bioinformatics is the solicitation of Information technology that analyses residing statistics and addresses biological problems. Integration of computers, software tools, databases, and techniques are endeavored to resolve organic problems. Genomics and Proteomics are the built-in working areas in Bioinformatics. R is the high-quality tool for examining organic data and committing to construct bioinformatics resolution effectively. Several equipment are handy for community comparison and visualization such as UniHi, String etc.

CRC is also called as bowel cancer. About 0.43% of total deaths are occurred due to CRC in Bangladesh [1]. Colon cancer develops first as colorectal polyps. Abnormal colon or rectum growth can turn into cancer. Around 9.6 million people die from cancer in Asia [2]. The goal is to achieve at least 60% survival for all children with cancer globally by 2030 [2].

Cancer is a frequent complex disease. Many genetic factors and genes have been pronounced to play an important role in its pathogenesis. Identification of genes that activate or accelerate the development of cancer has been one of the essential goals in cancer research [3]. Globally greater than 1 million people get affected by CRC yearly, resulting in about 0.5 million deaths [4].

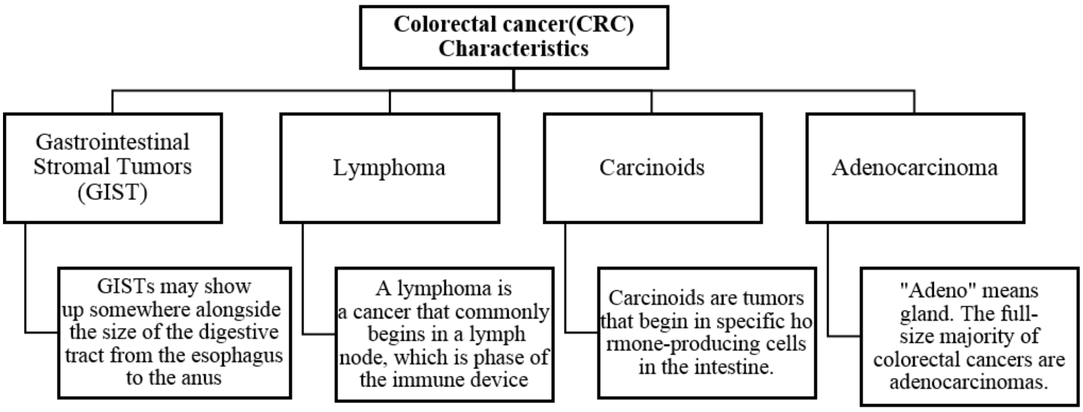

Colorectal cancer(CRC) Characteristics
Gastrointestinal Stromal Tumors (GIST)
Lymphoma
Carcinoids
Adenocarcinoma
GISTs may show up somewhere alongside the size of the digestive tract from the esophagus to the anus
A lymphoma is a cancer that commonly begins in a lymph node, which is phase of the immune device
Carcinoids are tumors that begin in specific hormone-producing cells in the intestine.
"Adeno" means gland. The full-size majority of colorectal cancers are adenocarcinomas.

Figure 1.1: Characteristics of CRC

CRC is the second and third leading cause of cancer mortality in men and women [5]. In 2008, 1.2 million people deaths were occurring for CRC characteristics worldwide [6].

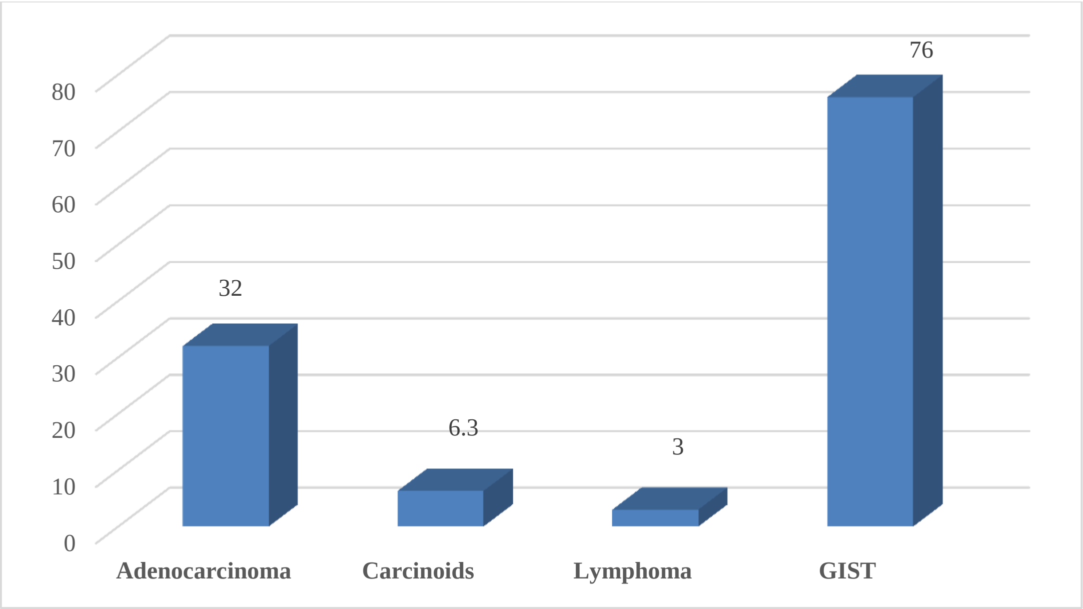


Figure 1.2: Impact rate (%) of CRC characteristics worldwide (2009-2017)

No one is aware of the precise causes of large intestine cancer. However, we tend to do apprehend that individuals with sure risk factors square measure a lot of seemingly than others to develop large intestine cancer. Studies have found the subsequent risk factors for large intestine cancer. Polyps have been divided into sessile and pedunculated according to theirs shape [7]. Most polyps begin (not cancer), however some polyps (adenomas) will become cancer. Adenocarcinoma, Lymphoma, Carcinoids, GIST are the common characteristics of CRC.

Adenocarcinoma is the worst killer of the CRC characteristics. The most frequent kind of bowel most cancers is called an adenocarcinoma, named after the gland cells in the lining of the bowel where the cancer first develops. It is a slow-growing cancer however can spread to the skull. The bowels are full bloody stool, rectal bleeding, stomach pain, and unexplained weight loss are the symptoms of Adenocarcinoma. It ultimately leads to a colon cancer. In 2016, there were 5375 deaths caused by bowel cancer in Australia. This represents the second highest number of cancer deaths in Australia [8]. There are now more than 1 million survivors of CRC in the United States [9].

Lymphoma is a group of blood cancers that boost from lymphocytes (a kind of white blood cell). The enlarged lymph nodes are normally painless. Fever, drenching sweats, un-intended weight loss, itching, and constantly feeling tired are the symptoms of Lymphoma CRC. Lymphoma is association with inflammatory bowel disease. A pathogenesis connection was suggested between CD and lymphoma. Lymphoma most often spreads to the lungs, liver, and brain. Worldwide, lymphomas developed in 566,000 people in 2012 and caused 305,000 deaths [10].

Carcinoid tumors are the most frequent malignant tumor of the appendix. Carcinoid is a slow growing cancer. They are most oftentimes associated with the small intestine.

Carcinoids can additionally be discovered in the rectum and stomach. They are known to develop in the liver. Adenomal pain, diarrhoea, nausea, rectal bleeding and rectal pain are the symptoms of Carcinoids. If it is all removed from an individual can be cured. In Australia, 24% patients were suffered with carcinoid tumor [11].

Gastrointestinal stromal tumors (GISTs) are the most common mesenchyme neoplasms of the gastrointestinal tract. GISTs may present with trouble swallowing, gastrointestinal bleeding or metastases (mainly in the liver). GIST is more common in men. GISTs can also take place somewhere alongside the size of the digestive tract from the oesophagus to the anus. GISTs occur in 10-20 per one million people. The majority of GISTs presents at ages between 50–70 years. The estimated incidence of GIST in the United States was approximately 5000 cases annually. Across most of the age spectrum, the incidence of GIST is similar in men and women [12].

1.2 Motivation

Unusualness in a gene is acknowledged as disease. A danger component is any attribute of eminent that rise the opportunity of developing ailment [13]. Risk factors for suicidal thoughts and conducts a 50-year research meta-analysis. Risk elements are interrelated. The energy of this interrelation is measured statistically. Adenomatous polyposis (FAP), Smoking, Diabetics, Inflammatory bowel disease (IBD) and Alcoholic are common risk factors of all CRC characteristics [14].

The outbreak of Smokers used to be discovered utmost in lower and middle-income countries. Smoking causes about 25% of all deaths in Bangladesh [19]. According to the World Health Statistics, cancer cases in Bangladesh have been estimated at 167 per 1, 00,000 population [20]. Cancer is the 6th leading cause of deaths and accounts for 10% of all mortality deaths in Bangladesh [21].

Type 2 diabetes mellitus is a long - term metabolic disorder with high blood sugar, insulin resistance and relative insulin shortage [22]. Type 2 diabetes comprises the majority of human beings with diabetes around the world. Generally it is the result of extra body weight and bodily inactivity. The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 2014 [23]. Its global prevalence was about 8% in 2011 and is predicted to rise to 10% by 2030 [24].

Inflammatory bowel disease (IBD) is an overcoat term that describes illnesses that involve chronic digestive tract inflammation. It has often been concept of as an autoimmune disease. Two predominant sorts of IBD are ulcerative colitis and Cohn's disease. In Europe (ulcerative colitis 505 per 100 000 in Norway; Cohn's disease 322 per 100 000 in Germany) and North America (ulcerative colitis) were the highest reported prevalence values (286 per 100 000 in the USA); (Cohn's disease 319 per 100 000 in Canada) [25].

A natural substance formed when a hydroxyl crew is substituted for a hydrogen atom in a hydrocarbon. The kind of alcohol used in alcoholic beverages, ethanol derives from fermenting sugar with yeast. The World Health Organization (WHO) estimates that there were around 3.3 million deaths worldwide per year due to use of alcohol [26].

1.3 Scope

This research analyses Adenocarcinoma, Lymphoma, Carcinoids, GIST whose are all characteristics of CRC. CRC is one of the most non-communicable cancers in the world. Diseases generated from frequent risk factors have an exceptional danger of genetic correlation at once or indirectly. KDD is the special strategy of information mining to discover a unique knowledge in large-scale data. KDD the data mining application along with Bioinformatics is contributing to human life. NCBI is the receptacle of biological data. Proteomics makes PPI more effective. Protein binds with different proteins by rules which turn into the favored chemical reactions of our body. Regulation is recognized in PPI which leads to finding key common genes along with frequent pathway among the proteins layout for all CRC characteristics.

1.4 Objectives

The principle targets of this thesis are given below:

- To find out genetic association among all characteristics of CRC.
- To find frequent genes among all CRC characteristics through Data mining with the usage of R.
- To generating PPI String tool is used.
- Identify key common genes based on PPI networks from all interaction sources.

1.5 Thesis Organization

This thesis document is enclosed with followings. The current chapter provides the introduction of this thesis. Chapter 2 describes related works. Research methodology is discussed in chapter 3. Chapter 4 shows experiments and corresponding results respectively. Finally, the conclusions are determined in section 5.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Bioinformatics creates a new imaginative and prescient in lifestyles sciences combining computer science, biology, statistics and mathematics. A protein is a complicated group of associated polypeptide chains linked by means of monovalent protein-protein interactions. Proteins hardly ever act on own as their functions have a tendency to be regulated with another one. PPI information and pathway data can be retrieved from STRING databases respectively as it is reliable bio tools for researchers [26]. This chapter is modeled to illustrate the background history of this study.

2.2 Related work

In [27], PPI used to predict heterogeneous sickness genes. Dataset was consist of four PPI based on 4 different species—Human, Drosophila melano-gaster, Caenorhabditis elegans, and Saccharomyces cerevisiae [27]. The collection was once 10,894 genes for 383 genetically heterogeneous hereditary diseases from 432 loci of candidate diseases. Protein-protein interplay set, Candidate gene prediction, Benchmark test, and Randomization take a look at techniques had been carried out for universal evaluation. 300 candidate genes were enumerated from 72, 940 protein-protein Interactions. Four diseases were found using R com-mon genes within genetically interrelated diseases. Gene data was gathered from the NCBI database. Genes have been checked using the database of ExPaSy. Using UniHi tool, PPI was evaluated.

PPI network regulatory interactions refine the basis for the identification of common genes as well as pathways. This research mainly focused on evaluating key common genes of CRC characteristics based on PPI networks.

2.3 Gene

Gene is the piece of DNA that includes genetic information. Basic physical and functional unit of functions of heredity is gene. Genes are made up of DNA. Genes (not all) act as the medium to make molecules known as protein. Each person holds two copies of each gene, one inherited from every parent. Most genes are similar in all human, less than 1% genes are exclusive amongst people.

Each gene has a special identify to keep track. Genes determine the entirety about us, from the outward bodily characteristics we can see to the in the back of the scenes constructions internal our cells that permit them to raise out all of our body functions [28].

2.4 Protein

Protein is the most important issue of all living organism. Protein is a molecule composed of polymers of amino acids joined collectively by way of peptide bonds. Proteins are large biomolecules or macromolecules. Proteins are made up of lengthy chains of amino acids which creates polypeptide chain polymer. When poly-peptide chain polymer folds a protein is created. Figure 2.1 is the view of protein structure.

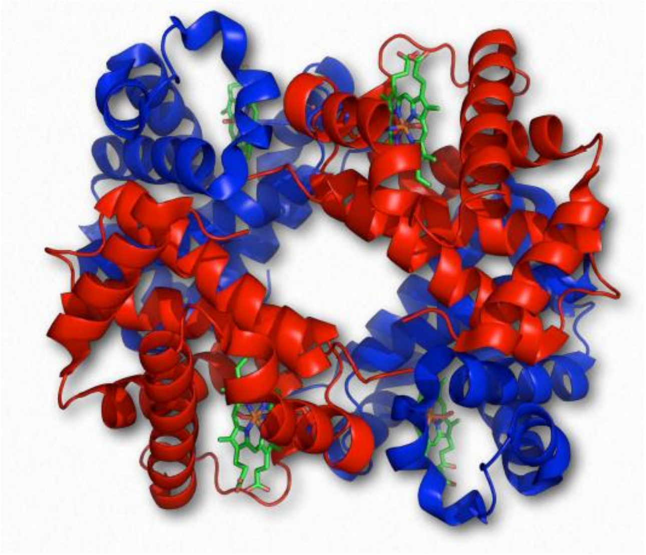


Figure 2.1: Structure of protein

The followings are the results of protein:

- Protein is the builder of our body such as bones, muscles, blood, teeth, finger etc.
- Cells of our physique are often hampered. Protein creates a new mobile on that hampered place.
- Protein generates heat for our body.
- Protein prevents our physique via developing antibody of the diseases.

So, it is clear that if our physique does no longer get proper proteins, there will take place many malfunctions in the body.

2.5 Proteomics

Proteomics is the large-scale study of proteins. Proteins are vital parts with many functions of living organisms. A proteome is a set of proteins generated in an organism, system, or biological context. Homo sapiens is an example of the proteome of species. The proteome is blended words of protein and genome and the naming was done by Marc Wilkins in 1994 at Macquarie University [29]. Also, Proteomics is the resolution of the entire protein complement of a cell, tissue, or organism under a tangible set of conditions [30].

Usages of proteomics are given below [31]:

- Look into protein appearance time.
- Inquire into fees of protein production, degradation, and steady-state, affluence.
- Investigate deportment of protein amongst subcellular compartments.
- Identify how a protein interacts with another.

2.6 Genomics

Genomics is the learning about complete genomes of organisms and accommodate elements from genetics. A genome is an organism's complete set of DNA, which include all of its genes. DNA recombinant combination, DNA sequencing methods, and bioinformatics are used to decide sequence, assemble and analyses the shape and characteristic of genomes in Genomics [32]. Genomics reason is at mixed characterization and quantification of genes which display protein manufacturing [33].

2.7 Protein-protein Interactions

PPI is the precise bodily connection between two or extra protein molecules which is performed via electrostatic forces together with the hydrophobic impact [34]. Proteins rarely tend to regulate alone or single. Molecular tactics are performed in a cell by using quite a number of protein factors organized by means of PPI. Aberrant PPI causes disease such as Alzheimer's diseases may also lead to cancer [35].

2.7.1 Advantages of Protein-protein Interactions

PPI helps to apprehend the consequences of interaction in a cell. Peptides are developed by way of PPI. In a single network activated or repressed proteins are decided through PPI. PPI leads the way to how locomotion of protein may additionally change. PPI is generated from tying proteins. The way of tying between proteins can alternate through dedication [36].

PPI information has been extracted from the Molecular Interaction database (MINT), the Biological General Repository for Interaction Datasets (BioGRID) and the Human Protein Reference Database (HPRD). The MINT database objectives are storing in structured format information about molecular interactions by means of extracting experimental details. The current model of MINT consists of over 240,000 experimentally confirmed PPI [37].

2.8 P-value

In statistical probability theory, for a given statistical model, the p - value or probability value or meaning is the probability that, if the null hypothesis is true, the statistical summary (such as the absolute sample mean difference between two comparison groups) would be greater than or equal to the actual results measured. The p - value is represented as the worst - case likelihood for a composite null hypothesis. The p - value ranges from 0 to 1 and is interpreted as follows:

- P - Values very near to the cutoff (0.05) are regarded as nominal
- A small p - value (typically 0.05) shows strong proof against the null hypothesis, so the null hypothesis is denied.
- A large p - value (> 0.05) shows weak evidence against the null hypothesis, so the null hypothesis is not dismissed.

2.9 Average clustering coefficient

When applied to a single node, the clustering coefficient is a measure of how complete a node's neighborhood is. It is the average clustering coefficient over all nodes in the network as applied to a whole network. A "small - world" influence may be stated by the clustering coefficient together with the mean shortest path. There are two versions of the measure: global and local. The global version was designed to provide an overall clustering indication in the network, while the local version provides an indication of single node embedding.

2.10 Interaction Score

The interaction with the highest score is selected when an agent becomes available and more than one interaction is going to wait. Interaction is something that happens when two or more objects affect each other. In the concept of interaction, the concept of a two - way effect is crucial as opposed to a one - way causal effect.

Each gene has a special identify to keep track. Genes determine the entirety about us, from the outward bodily characteristics we can see to the in the back of the scenes constructions internal our cells that permit them to rise out all of our body functions [28].

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

We are living in statistics age. Terabytes or petabytes information is pouring into the computer networks each and every day. A human physique is having approximately 19000-2000 protein coding genes. So, when it comes about tens of millions of people's organic data, Bioinformatics has a considerable amount of one-of-a-kind labelled data such DNA sequencing, protein-protein interaction etc. Continuously the growth of genomic and proteomic data is increasing. Obtaining specific outcomes from the sea of records needs a substantial procedure. Bioinformatics and data mining are developing as cross-functional science.

Data mining utility and strategies are active fields in Bioinformatics to solve biological problems. It has been found that cross-pollination between data mining and bioinformatics are very effective [25].

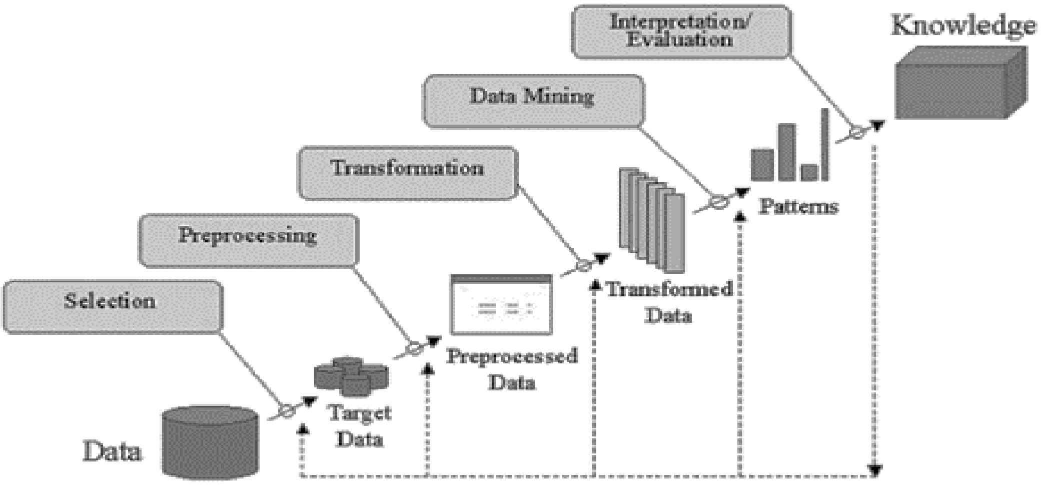


Figure 3.1: Steps of KDD process

Data mining and Knowledge Discovery in Database (KDD) are used interchangeably. KDD the facts mining approach is used to extract beneficial expertise from large-scale data. The following steps in Figure 3.1 are described below:

Data- Learn about the application domain, prior relevant works and what the goal is.

- **Selection**- Learn about the software domain, prior relevant works and what the intention is.
- **Preprocessing-** Cleanse information and cast off noise from data. Set techniques for handling missing data fields. Alteration of data as per requirements.
- **Transformation**- Reduction and projection of records are performed at this step. Simplify dataset via deleting undesired data. Set elements to exhibit facts depending on the purpose or task.
- **Data Mining**- Identify KDD goal with information mining strategies or algorithms to extract hidden sample and symbolize the pattern in a form or a set of such representations as classification guidelines or trees, regression, clustering etc.
- **Interpretation or Evaluation**- Understand quintessential knowledge from the mined patterns

Adopt the knowledge to function similarly moves and acquiring the conclusion.

3.2 Proposed Method

Protein-Protein Interaction is a sequential procedure. Various steps are carried out to obtain the PPI. Figure 3.2 delimitates graphical representation providing a stepwise clear realization of this thesis evaluation method. The steps are determined to extract key common genes. R language has used to complete the steps and attain the goal. String is used to obtain the PPI networks from mined common genes. Each step of the flowchart is accomplished in the following subsections from 3.2.1 to 3.2.7.

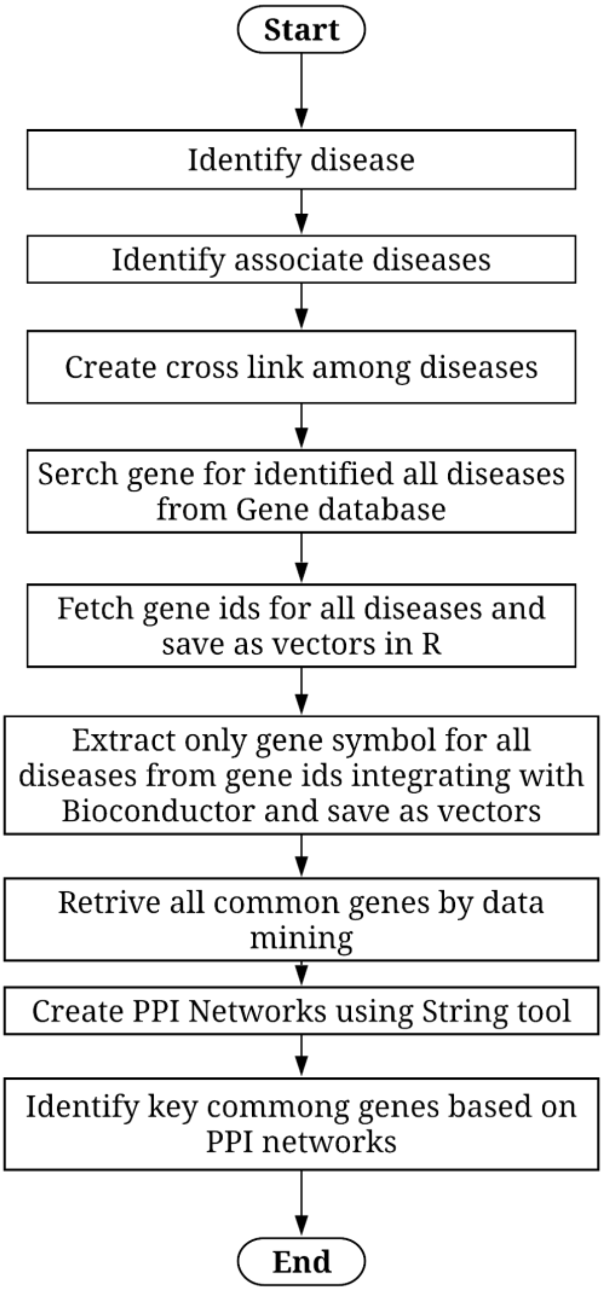

Start
Identify disease
Identify associate diseases
Create cross link among diseases
Serch gene for identified all diseases from Gene database
Fetch gene ids for all diseases and save as vectors in R
Extract only gene symbol for all diseases from gene ids integrating with Bioconductor and save as vectors
Retrive all common genes by data mining
Create PPI Networks using String tool
Identify key commong genes based on PPI networks
End

Figure 3.2: Flowchart of proposed methodology

3.2.1 Gene Search and collection

National Center for Biotechnology Information (NCBI) is referred to as the mother of Bioinformatics. It is a vital supply for bioinformatics tools and services as well as biological data. All biological information is saved right here divergently. Different categorized databases are attainable such as Gene, Nucleotide etc. The databases are available and downloadable through the internet or built-in tools. rEnterez is the package deal in R that receives get entry to in various NCBI databases including gene. For this thesis, genes are amassed associated with CRC characteristics. The pseudo code of the collection process is shown in Figure 3.3 respectively.

```
library(rentrez)
object_name <- entrez_search(db="database name", term="disease name")
object_name
object_name$ids
```

Figure 3.3: Gene collection process using R

3.2.2 Pre-Processing and Filtering

In prior step genes of each and every type are accrued for all organisms. Only Human genes are needed. This step collects genes only for the human organism or Homo sapiens. Here clearing noise from accumulated facts is known as pre-processing. So, gathered information is filtered and only human genes are added out. Just modify the pseudo code of the Figure 3.3.

```
library(rentrez)
object_name <- entrez_search(db="database name", term="(disease name) AND Homo Sapiens")
object _name
object_name$ids
```

Figure 3.4: Gene pre-processing and filtering

3.2.3 Cross-linkage Gene collection

Co-related genes are collected in this step. Total 11 combination are evaluated among all types of CRSs. Genes are collected among 2 types, 3 types and 4 types combinations. Genes are calculated. Pseudo code is given in the Figure 3.5.

```
object_name <- entrez_search(db="database name", term="(disease name AND disease name)"
object_name
object_name$ids
```

Figure 3.5: Cross linkage gene collection procedure using R

3.2.4 Gene Sorting

Gene data is consist of many elements such as aliases, gene symbol, gene id etc. To obtain aim gene image is needed. Gene symbol is the protein and aberrant PPI causes disease. Bio conductor an open supply software program platform is carried out for many bioinformatics strategies integrating R. Bio conductor is used to type gene image from gene data and saved in the database. Gene symbol for each kind is sorted. The Figure 3.6 shows the pseudo code of gene sorting.

```
source("https://bioconductor.org/biocLite.R")
biocLite("org.Hs.eg.db")
library(org.Hs.eg.db)
biocLite("annotate")
library(annotate)
object_name<- c(gene_ids)
lookUp(object_name, 'org.Hs.eg', 'SYMBOL')
```

Figure 3.6: Gene sorting procedure using R

3.2.5 Gene Mining

Data mining methods are used to pick out relevant data. This step is very crucial because any fault can miss out an important gene that turns error output. Further, a large amount of information can commit the result complicated. Read gene symbols of all characteristics and cross-linkages gene symbol in R and common genes are mined among them. The mined common genes are stored in a database and validated the usage of Expasy database to check whether gene symbols are correct or not. The mined genes are in contrast with the top 100 and 50 interrelated genes and store in a database. The Figure 3.7 depicts pseudo code of Gene mining.

```
#Create vectors for each types including all cross-linkages
object_name <- c(gene_symbols)
#Create an object for store common gene
object_name <- Reduce(intersect, list(object_names)
#Take 100 gene symbol from each type and Create a data frame
Object_name <- c(top 100 gene symbol)
Object_name <- data.frame(Object_name )
Object_name
#Read mined common gene symbols as a vector
Object_name <- c(common gene symbol)
#Create a vector for store common gene from top 100 gene symbol data frame
Object_name <- Reduce(intersect, list(object_name$column-name)
#Then create a vector take top 50 gene from data frame
Object_name <- head(object_name, 50)
#Create a vector for store common gene from top 50 gene symbols
Object_name <- Reduce(intersect, list(object_name$column-name)
#Create another data frame for store gene symol for common from all, common from top 100 and 50
Object_name <- data.frame(object_name)
```

Figure 3.7: Gene mining procedure using R

3.2.6 Create Protein-protein interactions network

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is web PPI visualization tool in Bioinformatics. It’s a superb tool for producing PPI. Protein interactions along with frequent pathways among co-related genes are considered in the PPI. From 6 co-related genes, PPI networks and common pathways are yielded using STRING tool. Thus, key common genes of CRC Characteristics are identified through measuring with different parameters based on all interaction sources provided in STRING (Version 11.0) tool.

3.3 Summary

At first, making use of the KDD steps CRC characteristics gene datasets are chosen from NCBI. Then the R language is used to accumulate cross-linkage genes. After mining, 6 mined common genes determined. Mined common genes verified by Expasy database. The pattern PPI is completed from mined common 6 genes using String tool.

CHAPTER 4

RESULTS AND DISCUSSION

This study mainly focused on finding key common genes among all characteristics of CRC. As accompanied the steps in the preceding chapters we obtained our goal. The genes responsible for CRC characteristics are loaded from the authentic database. Based on this representative sample, our whole finding was operated and carried out through sketching the PPI networks among diseases. KDD tactics are followed to consider liable genes and STRING is used to create PPI. For better understanding, this chapter is divided into following subsections.

.

4.2 Gene Optimization

Bioinformatics has a widespread amount of biological data. Optimization along with Bioinformatics makes the quality use of data. R language is used to accumulate gene datasets from the NCBI database. The following subsections 4.2.1 to 4.2.3 suggest how optimization was done. The Figure 4.1 also shows the format view of gene optimization.

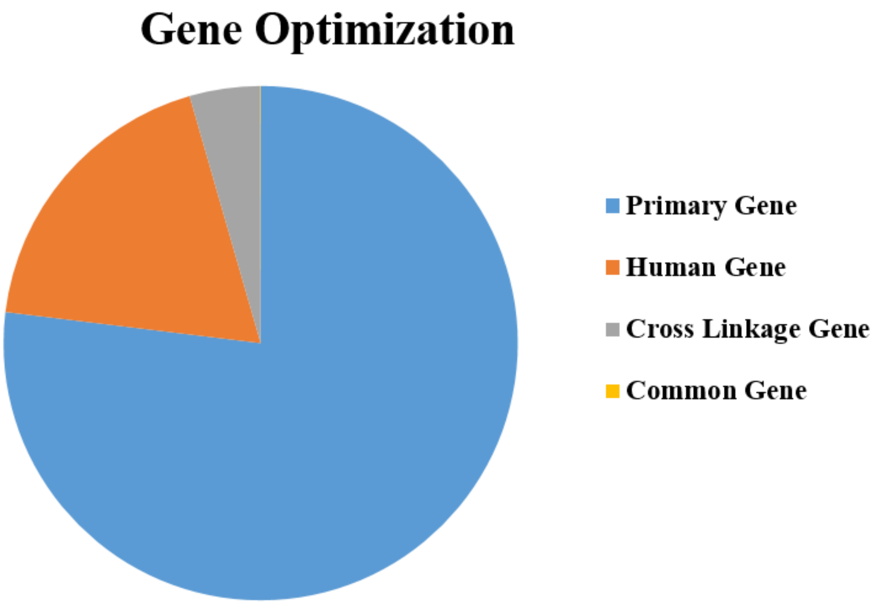


Figure 4.1: Scenario of gene optimization by pie chart

4.2.1 Gene collection, filter, preprocess and transform

The collection of genes are counted as Adenocarcinoma as 1853, Lymphoma as 126, Carcinoid as 58, GISTs as 66 and the total amount is 3263. The collected genes are accountable for all organisms and the amount is very large that makes the barrier to get genuine data. By filtering, only human genes are collected and the amount turned less to 3263.The Figure 4.2 shows the comparison between non-filter and filtered (Human) genes for all CRC characteristics.

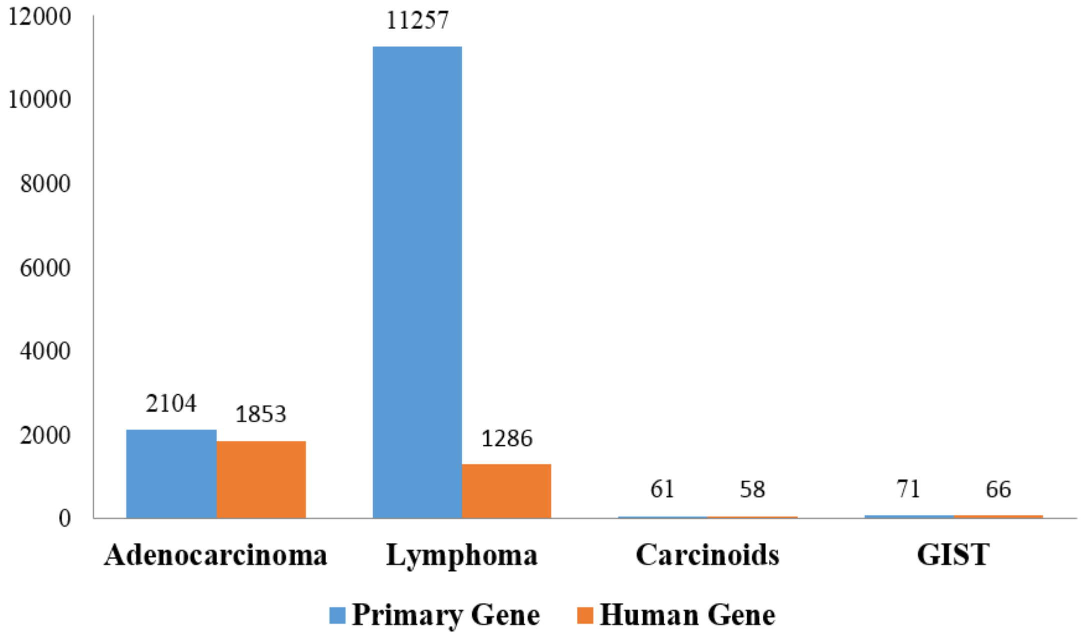


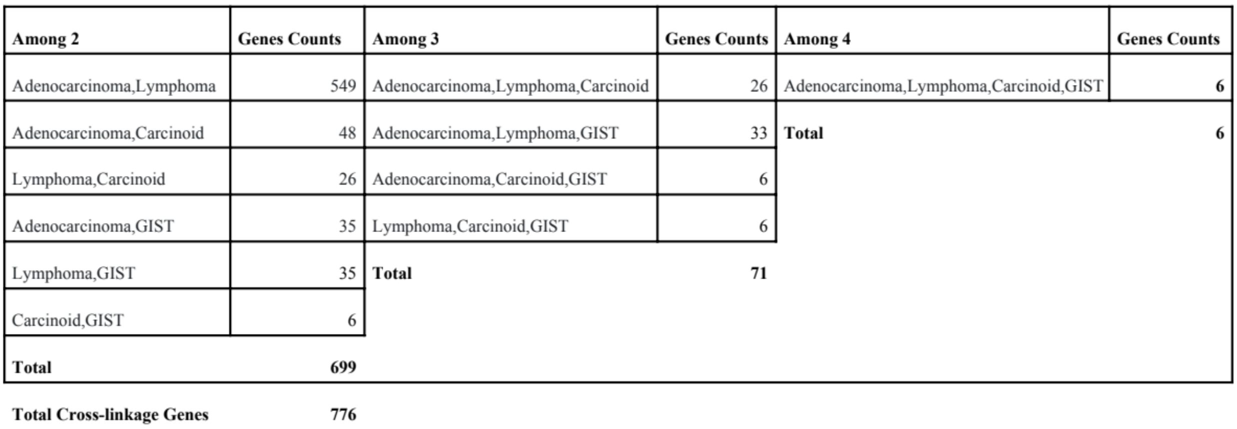
Figure 4.2: Difference between primary and filtered genes

By pre-process, genes are stored in separate databases for each characteristic. Gene information is consisting of many factors such as gene id, gene symbol, nucleotide sequence, aliases etc. For our study gene symbol is wanted as it is protein and protein binds with some other protein to run human body functions. So, from filtered gene information gene symbols are retrieved and saved in the database for further study.

4.2.2 Cross-Linkage gene collection

After amassing the genes, the cross-linkage was once utilized to located interrelated genes among Adenocarcinoma, Lymphoma, Carcinoid and GISTs. The cross connection among ailment helps to locate the genetic association. Cross-linkage combination was investigated among 2, 3 and 4 types. The Table 4.1 shows the calculation of cross-linkage genes among 4 CRC characteristics.

Table 4.1: Collected cross-linkage genes for CRC characteristics



Among 2	**Genes Counts**	**Among 3**	**Genes Counts**	**Among 4**	**Genes Counts**
Adenocarcinoma,Lymphoma	549	Adenocarcinoma,Lymphoma,Carcinoid	26	Adenocarcinoma,Lymphoma,Carcinoid,GIST	**6**
Adenocarcinoma,Carcinoid	48	Adenocarcinoma,Lymphoma,GIST	33	**Total**	**6**
Lymphoma,Carcinoid	26	Adenocarcinoma,Carcinoid,GIST	6		
Adenocarcinoma,GIST	35	Lymphoma,Carcinoid,GIST	6		
Lymphoma,GIST	35	**Total**	**71**		
Carcinoid,GIST	6				
Total	**699**				
Total Cross-linkage Genes	**776**				

4.2.3 Gene Mining

This is the fundamental step of gene optimization as well as this study. Corresponding genes for all characteristics are saved in databases. Cross-linkage genes are stored in the database too. Then interrelated cross-linkage genes are in contrast with all genes of CRC characteristics. From mining, 6 common genes had been found. Genes are EGFR, IGF1, MTOR, CD274, MKI67 and PDGFRA namely. These genes were verified with the usage of EXPASY database. Then common genes from the top 100 and top 50 are searched and stored comparing with 6 common genes.

4.3 PPI Network and common genes

Generating the PPI network of 6 mined common genes STRING tool is used. The PPI networks present the interaction pathways of genes based on their protein regulations. To identify significant result, the PPI networks are analyzed at four different levels (low, medium, high, and highest) provided in the STRING tool based on interaction score, p-value and co-efficiency. Figure 4.3 displays the PPI network among 6 common genes with low confidence interaction score, p-value and co-efficiency.

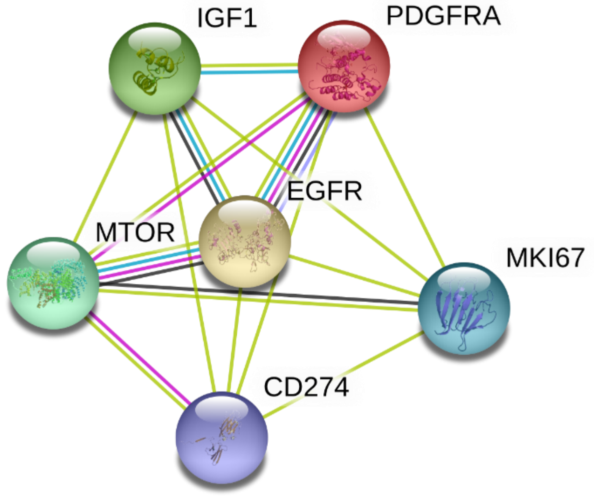


Figure 4.3: Protein-protein interaction network of CRC characteristics using STRING with low confidence

Continuously, Figure 4.4 and Figure 4.5 show the PPI networks with medium and high confidence interaction with the same parameters. From Figure 4.4 and Figure 4.5, it is clearly seen that 5 genes are interconnected with each other through protein interaction pathways in the PPI

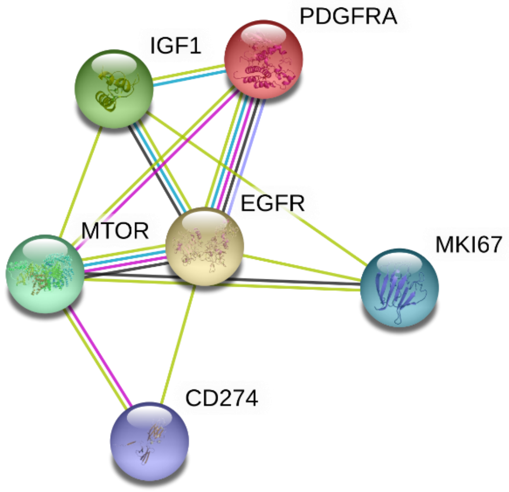


Figure 4.4: Protein-protein interaction network of CRC characteristics using STRING with medium confidence

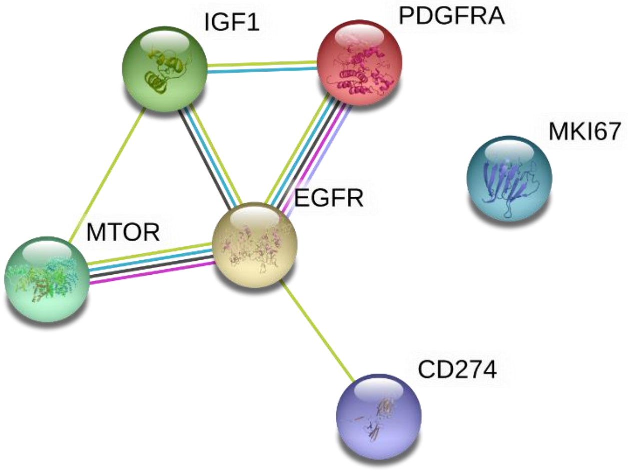


Figure 4.5: Protein-protein interaction network of CRC characteristics using STRING with high confidence

Figure 4.6 appears the PPI network with the highest confidence interaction score, p-value and co-efficiency. Visualizing in Figure 6 protein interaction pathways are available among 3 common genes in one cluster. CRC characteristics are affected with one another more precisely by these highest confident common genes through their regulation pathways.

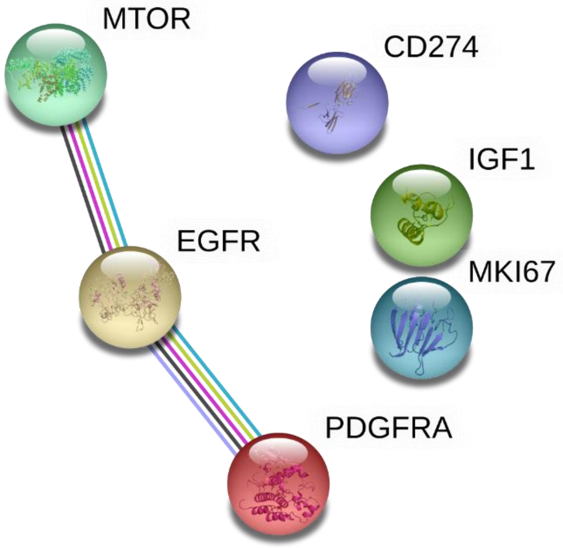


Figure 4.6: Protein-protein interaction network of CRC characteristics using STRING with highest confidence

Among the 6 common genes finally, highly 3 key common genes namely MTOR, EGFR and PDGFRA were found with the highest confidence level. Average local clustering coefficient, p-value, and interaction scores are shown in Table 4.2.

Table 4.2: Protein-protein interactions of 6 mined common genes from all active interaction sources in string 11.0

Confidence Level	Interaction Score	P-Value	Average Local Clustering Coefficient	Number of Common Genes
Highest Confidence	0.900	0.00509	0.333	3
High Confidence	0.700	0.0045	0.5	5
Medium Confidence	0.400	0.00109	0.667	5
Low Confidence	0.150	0.0415	0.833	6

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

The previous chapters are indispensable for resulted in this chapter. This chapter narrates findings along with contributions and the future expansion in the following sections based on overall performances and evaluations of preceding chapters.

5.1 Findings and Contributions

This study focuses on the threat evaluation and PPI networks creation of CRC characteristics, the disease that impact human life and top of all non-communicable illnesses in the world. Foremost Colorectal cancer is identical as CRC, in which 4 characteristics are responsible for the world deaths and various are standing at main positions. The contribution of proteomics and genomics in bioinformatics are incredible. The advancement of bioinformatics equipment and database explore a new research area and made future mission apparent. Protein-based therapeutics has additionally been developed in bio-informatics broadly.

Common danger elements are brought on among all CRC characteristics. KDD is carried out to extract frequent genes from NCBI gene dataset. Using String tool, PPI networks are generated from mined common genes. Interconnected key common genes are identified in the PPI community based on all interaction sources, average local clustering co-efficient, interaction score provided by String tool.

5.2 Recommendations for Future Works

Bioinformatics creates new fascinating research areas. Proteomics enhanced PPI network more effective. The future decision of this study is to work on various different correlated diseases to identify key common genes based on PPI.

References

[1] Parveen, R., Rahman, S. S., Sultana, S. A., & Habib, Z. H. (2015). Cancer types and treatment modalities in patients attending at Delta medical college hospital. Delta Medical College Journal, 3(2), 57-62.

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), 394-424.

[3] Susswein, L. R., Marshall, M. L., Nusbaum, R., Postula, K. J. V., Weissman, S. M., Yackowski, L., ... & Gibellini, F. (2016). Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. Genetics in Medicine, 18(8), 823.

[4] Deding, U., Torp-Pedersen, C., & Bøggild, H. (2018). The association between immigration status and ineligible stool samples for colorectal cancer screening. Cancer epidemiology, 57, 74-79.

[5] Dong, Z., Zheng, L., Liu, W., & Wang, C. (2018). Association of mRNA expression of TP53 and the TP53 codon 72 Arg/Pro gene polymorphism with colorectal cancer risk in Asian population: a bioinformatics analysis and meta-analysis. Cancer management and research, 10, 1341.

[6] Jemal, A., Center, M. M., DeSantis, C., & Ward, E. M. (2010). Global patterns of cancer incidence and mortality rates and trends. Cancer Epidemiology and Prevention Biomarkers, 19(8), 1893-1907.

[7] Mainprize, K. S., Mortensen, N. M., & Warren, B. F. (1998). Early colorectal cancer: recognition, classification and treatment. British journal of surgery, 85(4), 469-476.

[8] Lew, J. B., St John, D. J. B., Xu, X. M., Greuter, M. J., Caruana, M., Cenin, D. R., ... & Canfell, K. (2017). Long-term evaluation of benefits, harms, and cost-effectiveness of the National Bowel Cancer Screening Program in Australia: a modelling study. The Lancet Public Health, 2(7), e331-e340.

[9] M. Padovani and C. Oliani, "Chemotherapy in Rectal Cancer," Rectal Cancer,pp.2 15–220.

[10]"Lymphoma," Wikipedia, [Online]. Available:

https://en.wikipedia.org/wiki/Lymphoma. [Accessed: 10-Sep-2018].

[11] O'Rourke, M. G., Lancashire, R. P., & Vattoune, J. R. (1986). Carcinoid of the small intestine. Australian and New Zealand Journal of Surgery, 56(5), 405-408.

[12] Heinrich, M. C., Corless, C. L., Demetri, G. D., Blanke, C. D., Von Mehren, M., Joensuu, H., ... & Kiese, B. (2003). Kinase mutations and imatinib response in patients with metastatic gastrointestinal stromal tumor. Journal of clinical oncology, 21(23), 4342-4349.

[13] Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., ... & Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychological Bulletin, 143(2), 187.

[14] "Figure 2f from: Tan M, Kamaruddin K (2016) Redescription of the little-known grasshopper Willemsella (Acrididae, Hemiacridinae) from Peninsular Malaysia.Biodiversity Data Journal 4: e7775. https://doi.org/10.3897/BDJ.4.e7775."

[15] Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L., Albertsen, H., ... & Sargeant, L. (1991). Identification and characterization of the familial adenomatous polyposis coli gene. Cell, 66(3), 589-600.

[16] Galle, T. S., Juel, K., & Bülow, S. (1999). Causes of death in familial adenomatous polyposis. Scandinavian journal of gastroenterology, 34(8), 808-812.

[17] Munding, J., & Tannapfel, A. (2014). Epidemiology of Colorectal Adenomas and Histopathological Assessment of Endoscopic Specimens in the Colorectum. Visceral Medicine, 30(1), 10-16.

[18] Citarda, F., Tomaselli, G., Capocaccia, R., Barcherini, S., Crespi, M., & Italian Multicentre Study Group. (2001). Efficacy in standard clinical practice of colonoscopic polypectomy in reducing colorectal cancer incidence. Gut, 48(6), 812-815.

[19] Alam, D. S., Jha, P., Ramasundarahettige, C., Streatfield, P. K., Niessen, L. W., Chowdhury, M. A. H., ... & Evans, T. G. (2013). Smoking-attributable mortality in Bangladesh: proportional mortality study. Bulletin of the World Health Organization, 91, 757-764.

[20] Siegel, R., Naishadham, D., & Jemal, A. (2013). Cancer statistics, 2013. CA: a cancer journal for clinicians, 63(1), 11-30.

[21] Baade, P. D., Youlden, D. R., & Krnjacki, L. J. (2009). International epidemiology of prostate cancer: geographical distribution and secular trends. Molecular nutrition & food research, 53(2), 171-184.

[22]"Diabetes mellitus type 2," Wikipedia, [Online]. Available:

https://en.wikipedia.org/wiki/Diabetes_mellitus_type_2.

[Accessed: 13-Sep-2018]..

[23] "Diabetes," World Health Organization, 07-Sep-2018. [Online].

Available: https://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: 15-Sep-2018].

[24] Akter, S., Rahman, M. M., Abe, S. K., & Sultana, P. (2014). Prevalence of diabetes and prediabetes and their risk factors among Bangladeshi adults: a nationwide survey. Bulletin of the World Health Organization, 92, 204-213A.

[25] Kuhs, H., & Eikelmann, B. (1988). Suspension of neuroleptic therapy in acute schizophrenia. Pharmacopsychiatry, 21(04), 197-202.

[26] Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., ... & Calderwood, M. A. (2019). Network-based prediction of protein interactions. Nature communications, 10(1), 1240.

[27] Cowen, L., Ideker, T., Raphael, B. J., & Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics, 18(9), 551.

[28]“Gene,” Wikipedia, [Online].

Available: https://en.wikipedia.org/wiki/Gene. [Accessed: 17-Sep-2018].

[29]“Proteomics,” Wikipedia, [Online].

Available:https://en.wikipedia.org/wiki/Proteomics. [Accessed: 20-Sep-2018].

[30] Mosa, K. A., Gairola, S., Jamdade, R., El-Keblawy, A., Al Shaer, K. I., Al Harthi, E. K., ... & Mahmoud, T. (2018). The promise of molecular and genomic techniques for biodiversity research and DNA barcoding of the Arabian Peninsula flora. Frontiers in plant science, 9.

[31] “What is proteomics?,” EMBL-EBI Train online, [Online].

Available: https://www.ebi.ac.uk/training/online/course/proteomics-introduction-ebi-resources/what-proteomics. [Accessed: 23-Sep-2018].

[32]“Genomics,” Wikipedia, [Online].

Available: https://en.wikipedia.org/wiki/Genomics.

[Accessed: 27-Sep-2018].

[33]“What is genomics?,” EMBL-EBI Train online, [Online].

Available:https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/what-genomics. [Accessed: 05-Oct-2018].

[34] Pattin, K. A., & Moore, J. H. (2009). Role for protein–protein interaction databases in human genetics. Expert review of proteomics, 6(6), 647-659.

[35] “Protein–protein interaction,” Wikipedia, [Online].

Available: https://en.wikipedia.org/wiki/Protein–protein_interaction. [Accessed: 10-Oct-2018].

[36]“Protein quality,” Wikipedia, [Online].

Available: https://en.wikipedia.org/wiki/Protein_quality.

[Accessed: 20-Oct-2018].

[37] Luo, T., Wu, S., Shen, X., & Li, L. (2013). Network cluster analysis of protein–protein interaction network identified biomarker for early onset colorectal cancer. Molecular biology reports, 40(12), 6561-6568.

[38] Raza, K. (2012). Application of data mining in bioinformatics. arXiv preprint arXiv:1205.1125.

Appendix – A

List of Abbreviations

CRC = Colorectal Cancer

WHO = World Health Organization

FAP = Familial Adenomatous Polyposis

GISTs = Gastrointestinal Stromal Tumor

KDD = Knowledge Discovery in Database

PPI = Protein-Protein Interaction

NCBI = National Information Centre of Biotechnology