Mining of DNA repair genomic and network data to elucidate regulatory processes involved in human diseases

SUBMITTED IN PARTIAL FULFILLMENT FOR THE REQUIREMENT OF
BACHELOR OF TECHNOLOGY
IN
BIO-INFORMATICS

By
Lokesh Sharma

Under the guidance of

Dr. Tiratha Raj Singh
Assistant Professor (Senior Grade)

Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
P.O.Waknaghat-173234
Himachal Pradesh (INDIA)

CONTENTS

List of Figures

List of Tables

DECLARATION

I hereby declare that the project titled "**Mining of DNA repair genomic and network data to elucidate regulatory processes involved in human diseases"** is submitted as a Project Work has been carried out by me at Jaypee University of Information Technology, Solan under the guidance of Dr. Tiratha Raj Singh**.** Any further extension, continuation or use of this project has to be undertaken with prior express written consent from the Supervisor, Jaypee University of Information Technology, Solan-173234.

I further declare that the project work or any part thereof has not been previously submitted for any degree or diploma in any university.

Signature: Name:

Date:

CERTIFICATE

This is to certify that the work entitled “**Mining of DNA repair genomic and network data to elucidate regulatory processes involved in human diseases”** submitted by “**Lokesh Sharma**” in partial fulfillment for the award of Degree of Bachelor of Technology in **Bioinformatics** of Jaypee University of Information Technology, Waknaghat 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.

Name of Supervisor Dr. Tiratha Raj Singh

Signature of Supervisor ………………………

Designation Assistant Professor (Senior Grade)

Date ………………………

ACKNOWLEDGEMENT

I acknowledge my sincere thanks to **Dr. Tiratha Raj Singh**, *Assistant Professor (Senior Grade), Jaypee University of Information Technology, JUIT, Solan* for giving me this great opportunity to have project work under him and throughout this project, enlightening me on various topics and creating a congenial environment for my work. I am sure it would continue to raise research interests in young students like me.

I am grateful to the Ashwani Kumar, PhD student in the department of BT and BI, Jaypee University of Information Technology, JUIT, who helped me throughout this project and supported me in this Project.

Date: ……….   Signature:

Summary

DNA repair is a collection of processes by which a cell identifies and corrects damage to the DNA molecules that encode its genome. The main focus of this work is to mine the DNA repair genomic data i.e. to find out the potential genes that play a significant role in repair mechanism. Genes uses different mechanism to cure the damage occurred in DNA. Based upon literature survey and available information, we selected 343 DNA repair genes which are potentially significant for repair mechanism in Homo Sapiens along with other relevant information such as cellular component, molecular function and biological processes. Other necessary information for all the 343 genes were computed and decision tree were generated in Rapid Miner. Based upon this mined information association rules were generated using the Weka tool. Similar approach was applied on two diseases i.e. colorectal cancer and endometrial cancer, datasets for which consists of 226 and 85 genes respectively. For these two diseases GO term was computed along with the cellular component, molecular function and biological process information and decision tree was made using GO term in Rapid Miner and the association rule using Weka tool for both the diseases. It has been observed that specificities regarding diseases in the study were followed similar to general approach and important entities such as genes and proteins were identified which could be plausible targets for therapeutic studies.

1.Introduction

1.1 DNA Repair

Our cells are constantly exposed to abuses from endogenous and exogenous agents that can introduce damage into our DNA and generate genomic instability. Many of these lesions cause structural damage to DNA and can alter or eliminate fundamental cellular processes, such as DNA replication or transcription. DNA lesions commonly include base and sugar modifications, single- and double-strand breaks, DNA-protein cross-links, and base-free sites. To counteract the harmful effects of DNA damage, cells have developed a specialized DNA repair system, which can be subdivided into several distinct mechanisms based on the type of DNA lesion. These processes include base excision repair, mismatch repair, nucleotide excision repair, and double-strand break repair, which comprise both homologous recombination and non-homologous end-joining. Although a complex set of cellular responses are elicited following DNA damage, this chapter provides an introduction to the specific molecular mechanisms of recognition, removal, and repair of DNA damage.

DNA repair is a collection of processes by which a cell identifies and corrects damage to the DNA molecules that encode its genome. In human cells, both normal metabolic activities and environmental factors such as UV light and radiation can cause DNA damage, resulting in as many as 1 million individual molecular lesions per cell per day. The DNA repair process is constantly active as it responds to damage in the DNA structure. When normal repair processes fail, and when cellular apoptosis does not occur, irrepairable DNA damage may occur, including double-strand breaks and DNA cross linkages[2].

The rate of DNA repair is dependent on many factors, including the cell type, the age of the cell, and the extracellular environment. A cell that has accumulated a large amount of DNA damage, or one that no longer effectively repairs damage incurred to its DNA. The DNA repair ability of a cell is vital to the integrity of its genome and thus to the normal functionality of that organism. Many genes that were initially shown to influence life span have turned out to be involved in DNA damage repair and protection.[3]

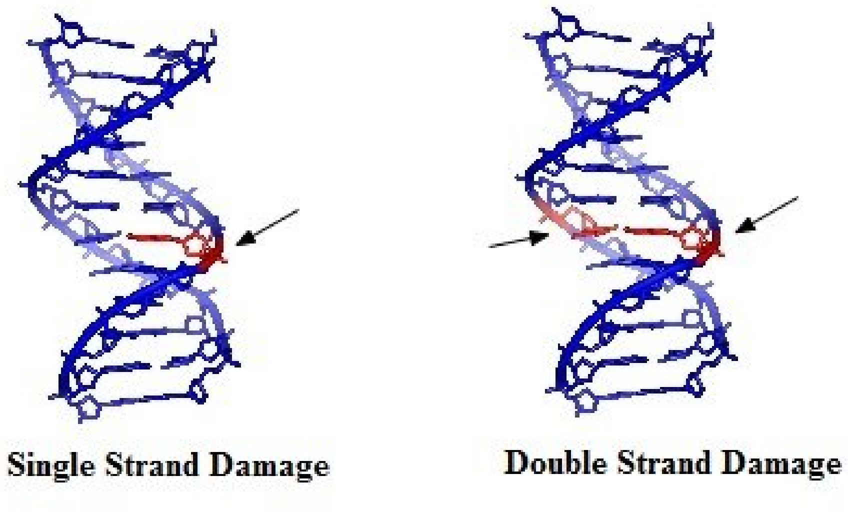

Fig. 1. Representation of single strand and double strand DNA damages. Source: - http://en.wikipedia.org/wiki/DNA_repair

Causes for DNA Damage

The various factors responsible for the damage of DNA can be subdivided into two main categories:-

Endogenous Damage, including replication errors and attack by reactive oxygen species produced from normal metabolic byproducts (spontaneous mutation), especially the process of oxidative deamination.

Exogenous Damage, caused by external agents like Ultraviolet [UV 200-400 nm] radiation from the sun. Other Radiation frequencies like X-rays and Gamma rays ,Hydrolysis or thermal disruption Certain plant toxins Human-made mutagenic chemicals, especially aromatic compounds that act as DNA intercalating agents viruses.

Different mechanisms of DNA Repair

Base Excision Repair (BER)

BER plays a major role in removing small, non-helix-distorting base lesions from the genome that could otherwise cause mutations by mispairing or lead to breaks in DNA during replication. BER is initiated by DNA glycosylases, which recognizes and removes specific damaged or inappropriate bases, forming AP sites. These are then cleaved by an AP endonuclease . The resulting single-strand break can then be processed by the process of BER.

The various base lesions repaired by BER includes:

Oxidized bases: 8-oxoguanine, 2,6-diamino-4-hydroxy-5-formamidopyrimidine (FapyG, FapyA)

Alkylated bases: 3-methyladenine, 7-methylguanine

Deaminated bases: hypoxanthine formed from deamination of adenine. Xanthine formed from deamination of guanine

Uracil inappropriately incorporated in DNA or formed by deamination of cytosine

Nucleotide Excision Repair (NER)

NER is an important mechanism by which the cell can prevent unwanted mutations by removing the vast majority of UV-induced DNA damage (mainly thymine dimers and 6-4-photoproducts). This mechanism mostly repairs bulky helix-distorting lesions as these NER enzymes recognize bulky distortions in the shape of the DNA double helix. Recognition of these distortions leads to the removal of a short single-stranded DNA segment that includes the lesion, creating a single-strand gap in the DNA, which is subsequently filled in by DNA polymerase, which uses the undamaged strand as a template for synthesis[1].

There are 9 major proteins involved in NER in mammalian cells and their names come from the diseases associated with the deficiencies in those proteins. XPA, XPB, XPC, XPD, XPE, XPF, and XPG all derive from Xeroderma pigmentosum and CSA and CSB represent proteins linked to Cockayne syndrome . Additionally, the proteins ERCC1, RPA, RAD23A, RAD23B, and others also participate in nucleotide excision repair. The importance of this repair mechanism is evidenced by the severe human diseases that result from in-born genetic mutations of NER proteins including Xeroderma pigmentosum and Cockayne's syndrome[1].

Mismatch Repair (MMR)

MMR recognizes and repairs erroneous insertion, deletion and mis-incorporation of bases that can arise during DNA replication and recombination, as well as repairs some forms of DNA damage.This process of repair is strand-specific and the mismatched bases including G/T or A/C pairing and other mismatches are commonly due to tautomerization of bases during synthesis.The damage is repaired by recognition of the deformity caused by the mismatch, determining the template and non-template strand, and excising the wrongly incorporated base and replacing it with the correct nucleotide.

There are a number of important mismatch repair proteins but three of these proteins are essential in detecting the mismatch and directing repair machinery to it- MutS, MutH and MutL. Mutations in the human homologues of the Mut proteins affect genomic stability, which can result in Microsatellite instability (MI)[1].

Homologous Recombination Repair (HRR)

Homologous recombination is mainly a genetic recombination in which nucleotide sequences are exchanged between two similar or identical molecules of DNA. This mechanism is mostly used by the cells to accurately repair harmful breaks that occur on both strands of DNA, known as double-strand breaks that are caused by ionizing radiation or DNA-damaging chemicals. If these breaks remain unrepaired, this can cause large-scale rearrangement of chromosomes in somatic cells which may in turn lead to Cancer.

Thus, recombination provides critical support for DNA replication in the recovery of stalled or broken replication forks, contributing to tolerance of DNA damage. There are many HRR proteins like Rad51, BRCA1 which play important role in the repair mechanisms and abberations in these genes have been implicated in a number of disorders including Genomic instability and contributes to cancer etiology[1].

Non-Homologous End Joining (NHEJ)

Non-homologous end joining (NHEJ) pathway mainly repairs the double-strand breaks (DSB's) in DNA, which if not repaired or misrepaired, can result in mutations, chromosome rearrangements and eventually in cell death. This pathway is named as "non-homologous" since the broken ends are ligated directly without the presence of a homologous template whereas in HRR, a homologous sequence is required to guide the repair. In mammals DSB's are primarily repaired by NHEJ and HRR, while HRR is mainly found in yeast and NHEJ can repair almost all kind of cells.

Deficiencies in DSB repair are associated with hereditary diseases such as Nijmegen breakage syndrome, Ataxia telangiectasia, Bloom’s syndrome and Breast cancer , etc. Several human syndromes are associated with the malfunctioning of NHEJ pathway[1].

1.2 Decision Tree

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

- **Information Gain**

Information gain is an impurity-based criterion that uses the entropy measure as the impurity measure.

$$InformationGain(ai, S) =$$

$$Entropy(y, S) - \sum_{vi,j \in dom(ai)} \frac{|\sigma ai = vi.jS|}{|S|}.Entropy(y, \sigma ai = vi, jS)$$

Where:

$$Entropy(y, S) = \sum_{cj \in dom(y)} \frac{-|\sigma y = cjS|}{|S|}.log2\frac{|\sigma y = cjS|}{|S|}$$

- **Gain Ratio**
  The gain ratio “normalizes” the information gain as follows

$$GainRatio(ai, S) = \frac{InformationGain(ai, S)}{Entropy(ai, S)}$$

- **ID3**

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan[4] used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains. The ID3 algorithm is considered as a very simple decision tree algorithm ID3 uses information gain as splitting criteria. The growing stops when all instances belong to a single value of target feature or when best information gain is not greater than zero. ID3 does not apply any pruning procedures nor does it handle numeric attributes or missing values. ID3 does not guarantee an optimal solution; it can get stuck in local optimums. It uses a greedy approach by selecting the best attribute to split the dataset on each iteration. One improvement that can be made on the algorithm can be to use backtracking during the search for the optimal decision tree.

ID3 can overfit to the training data, to avoid over fitting, smaller decision trees should be preferred over larger ones. This algorithm usually produces small trees, but it does not always produce the smallest possible tree. ID3 is harder to use on continuous data. If the values of any given attribute is continuous, then there are many more places to split the data on this attribute, and searching for the best value to split by can be time consuming. The ID3 algorithm is used by training on a dataset S to produce a decision tree which is stored in memory. At runtime, this decision tree is used to classify new unseen test cases by working down the decision tree using the values of this test case to arrive at a terminal node that tells you what class this test case belongs to.

- **C4.5**

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan[5]. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.It uses gain ratio as splitting criteria. The splitting ceases when the number of instances to be split is below a certain threshold. Error–based pruning is performed after the growing phase. C4.5 can handle numeric attributes. It can induce from a training set that incorporates missing values by using corrected gain ratio criteria. C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S = {s_*1*, s_*2*, ...} of already classified samples. Each sample si consists of a p-dimensional vector (x_{*1*,i}, x_{*2*,i}, ...,x_{p,i}) , where the xj represent attributes or features of the sample, as well as the class in which si falls. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists. J48 is an open source Java implementation of the C4.5 algorithm in the weka data mining tool.

2. Diseases

2.1 Colorectal Cancer

Colorectal cancer (also known as colon cancer, rectal cancer or bowel cancer) is the development of cancer in 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. Signs and symptoms may include blood in the stool, a change in bowel movements, weight loss, and feeling tired all the time. Risk factors for colorectal cancer include lifestyle, older age, and inherited genetic disorders that only occur in a small fraction of the population. Other risk factors include diet, smoking, alcohol, lack of physical activity, family history of colon cancer and colon polyps, presence of colon polyps, race, exposure to radiation, and even other diseases such as diabetes and obesity. A diet high in red, processed meat, while low in fiber increases the risk of colorectal cancer [9].

Treatments used for colorectal cancer may include some combination of surgery, radiation therapy, chemotherapy and targeted therapy. Cancers that are confined within the wall of the colon may be curable with surgery while cancer that has spread widely are usually not curable, with management focusing on improving quality of life and symptoms. Five year survival rates in the United States are around 65%. This, however, depends on how advanced the cancer is, whether or not all the cancer can be removed with surgery, and the person's overall health. Globally, colorectal cancer is the third most common type of cancer making up about 10% of all cases. In 2012 there were 1.4 million new cases and 694,000 deaths from the disease. It is more common in developed countries, where more than 65% of cases are found. It is less common in women than men [6].

The signs and symptoms of colorectal cancer depend on the location of the tumor in the bowel, and whether it has spread elsewhere in the body (metastasis). The classic warning signs include: worsening constipation, blood in the stool, decrease in stool caliber (thickness), loss of appetite, loss of weight, and nausea or vomiting in someone over 50 years old. While rectal bleeding or anemia are high-risk features in those over the age of 50, other commonly-described symptoms including weight loss and change in bowel habit are typically only concerning if associated with bleeding.

Greater than 75-95% of colon cancer occurs in people with little or no genetic risk. Other risk factors include older age, male gender, high intake of fat, alcohol or red meat, obesity, smoking, and a lack of physical exercise. Approximately 10% of cases are linked to insufficient activity. The risk

for alcohol appears to increase at greater than one drink per day. Drinking 5 glasses of water a day is linked to a decrease in the risk of colorectal cancer and adenomatous polyps.
Diagnosis of colorectal cancer is via sampling of areas of the colon suspicious for possible tumor development typically done during colonoscopy or sigmoidoscopy, depending on the location of the lesion. The extent of the disease is then usually determined by a CT scan of the chest, abdomen and pelvis. There are other potential imaging test such as PET and MRI which may be used in certain cases. Colon cancer staging is done next and based on the TNM system which is determined by how much the initial tumor has spread, if and where lymph nodes are involved, and the extent of metastatic disease.

The microscopic cellular characteristics of the tumor are usually reported from the analysis of tissue taken from a biopsy or surgery. A pathology report will usually contain a description of cell type and grade. The most common colon cancer cell type is adenocarcinoma which accounts for 98% of cases. Other, rarer types include lymphoma and squamous cell carcinoma [23].

2.2 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 40% of cases are related to obesity. Endometrial cancer is also associated with excessive estrogen exposure, high blood pressure and diabetes. Whereas taking estrogen alone increases the risk of endometrial cancer, taking both estrogen and progesterone in combination, as in most birth control pills, decreases the risk. Between two and five percent of cases are related to genes inherited from the parents. Endometrial cancer is sometimes loosely referred to as "uterine cancer", although it is distinct from other forms of uterine cancer such as cervical cancer, uterine sarcoma, and trophoblastic disease [24].

The most frequent type of endometrial cancer is endometrioid carcinoma, which accounts for more than 80% of cases. Endometrial cancer is commonly diagnosed by endometrial biopsy or by taking samples during a procedure known as dilation and curettage. A pap smear is not typically sufficient to show endometrial cancer. Regular screening in those at normal risk is not called for.

The leading treatment option for endometrial cancer is abdominal hysterectomy (the total removal by surgery of the uterus), together with removal of the fallopian tubes and ovaries on both sides, called a bilateral salpingo-oophorectomy. In more advanced cases, radiation therapy, chemotherapy or hormone therapy may also be recommended. If the disease is diagnosed at an early stage, the outcome is favourable, and the overall five-year survival rate in the United States is greater than 80%[8].

There are several types of endometrial cancer, including the most common endometrial carcinomas, which are divided into Type I and Type II subtypes. There are also rarer types including endometrioid adenocarcinoma, uterine papillary serous carcinoma, and uterine clear-cell carcinoma. Vaginal bleeding or spotting in women after menopause occurs in 90% of endometrial cancer. Bleeding is especially common with adenocarcinoma, occurring in two-thirds of all cases. Abnormal menstrual cycles or extremely long, heavy, or frequent episodes of bleeding in women before menopause may also be a sign of endometrial cancer. Symptoms other than bleeding are not common. Other symptoms include thin white or clear vaginal discharge in postmenopausal women. More advanced disease shows more obvious symptoms or signs [24] that can be detected on a physical examination. The uterus may become enlarged or the cancer may spread, causing lower abdominal pain or pelvic cramping. Painful sexual intercourse or painful or difficult urination are less common signs of endometrial cancer. The uterus may also fill with pus (pyometrea). Of women with these less common symptoms (vaginal discharge, pelvic pain, and pus), 10–15% have cancer.

Risk factors for endometrial cancer include obesity, diabetes mellitus, breast cancer, use of tamoxifen, never having had a child, late menopause, high levels of estrogen, and increasing age. Immigration studies (migration studies), which examine the change in cancer risk in populations moving between countries with different rates of cancer, show that there is some environmental component to endometrial cancer. These environmental risk factors are not well characterized. Most of the risk factors for endometrial cancer involve high levels of estrogens. An estimated 40% of cases are thought to be related to obesity. In obesity, the excess of adipose tissue increases conversion of androstenedione into estrone, an estrogen. Higher levels of estrone in the blood causes less or no ovulation and exposes the endometrium to continuously high levels of estrogens. Obesity also causes less estrogen to be removed from the blood. Polycystic ovary syndrome (PCOS), which also causes irregular or no ovulation, is associated with higher rates of endometrial cancer for the same reasons as obesity. Specifically, obesity, type II diabetes, and insulin resistance are risk factors for Type I endometrial cancer. Obesity increases the risk for endometrial cancer by 300–400%.

Estrogen replacement therapy during menopause when not balanced (or "opposed") with progestin is another risk factor [10].

Higher doses or longer periods of estrogen therapy have higher risks of endometrial cancer. Women of lower weight are at greater risk from unopposed estrogen. A longer period of fertility—either from an early first menstrual period or late menopause—is also a risk factor. Unopposed estrogen raises an individual's risk of endometrial cancer by 2–10 fold, depending on weight and length of therapy. In trans men who take testosterone and have not had a hysterectomy, the conversion of testosterone into estrogen via androstenedione may lead to a higher risk of endometrial cancer.

Smoking and the use of progestin are both protective against endometrial cancer. Smoking provides protection by altering the metabolism of estrogen and promoting weight loss and early menopause. This protective effect lasts long after smoking is stopped. Progestin is present in the combined oral contraceptive pill and the hormonal intrauterine device (IUD). Combined oral contraceptives reduce risk more the longer they are taken: by 56% after four years, 67% after eight years, and 72% after twelve years. This risk reduction continues for at least fifteen years after contraceptive use has been stopped. Obese women may need higher doses of progestin to be protected. Having had more than five infants (grand multiparity) is also a protective factor, and having at least one child reduces the risk by 35%. Breastfeeding for more than 18 months reduces risk by 23%. Increased physical activity reduces an individual's risk by 38–46%. There is preliminary evidence that consumption of soy is protective [15].

Diagnosis of endometrial cancer is made first by a physical examination and dilation and curettage (removal of endometrial tissue; D&C). This tissue is then examined histologically for characteristics of cancer. If cancer is found, medical imaging may be done to see whether the cancer has spread or invaded tissue.

The primary treatment for endometrial cancer is surgery; 90% of women with endometrial cancer are treated with some form of surgery. Surgical treatment typically consists of hysterectomy including a bilateral salpingo-oophorectomy, which is the removal of the uterus, and both ovaries and Fallopian tubes. Lymphadenectomy, or removal of pelvic and para-aortic lymph nodes, is performed for tumors of histologic grade II or above. Lymphadenectomy is routinely performed for all stages of endometrial cancer in the United States, but in the United Kingdom, the lymph nodes are typically only removed with disease of stage II or greater. The topic of lymphadenectomy and what survival benefit it offers in stage I disease is still being debated. In stage III and IV cancers, cytoreductive surgery is the norm, and a biopsy of the omentum may also be included. In stage IV disease, where there are distant metastases, surgery can be used as part of palliative therapy. Laparotomy, an open-abdomen procedure, is the traditional surgical procedure; however, laparoscopy (keyhole surgery) is associated with lower operative morbidity. The two procedures have no difference in overall survival. Removal of the uterus via the abdomen is recommended over removal of the uterus via the vagina because it gives the opportunity to examine and obtain washings of the abdominal cavity to detect any further evidence of cancer. Staging of the cancer is done during the surgery.

The few contraindications to surgery include inoperable tumor, massive obesity, a particularly high-risk operation, or a desire to preserve fertility. These contraindications happen in about 5–10% of cases. Women who wish to preserve their fertility and have low-grade stage I cancer can be treated with progestins, with or without concurrent tamoxifen therapy. This therapy can be continued until the cancer does not respond to treatment or until childbearing is done. Uterine perforation may occur during a D&C or an endometrial biopsy. Side effects of surgery to remove endometrial cancer can specifically include sexual dysfunction, temporary incontinence, and lymphedema, along with more common side effects of any surgery, including constipation [24].

3. Objective

Mining of DNA repair genomic and network data to elucidate regulatory processes involved in human diseases

DNA damage can be induced by a large number of physical and chemical agents from the environment as well as compounds produced by cellular metabolism. This type of damage can interfere with cellular processes such as replication and transcription, resulting in cell death and/or mutations. We have huge data related to genes with respect to DNA repair mechanism in various forms, individually at different places by different scientists and scholars. In this Project, I aim to put my efforts to first gather all those type of possible data, convert it into information through various kind of computational and statistical analyses, and finally to produce some biologically meaningful information as knowledge.

4. Methodology

Various standard resources and databases like NCBI (National Center for Biotechnology Information) were scanned which contain information about genes/proteins and searched for DNA repair, and collected information about each gene in the databases and also considered many latest journals for the selection of new potential DNA repair genes which are playing major role in Repair Mechanisms in Homo sapiens and also searched for Gene Ontology information. Decision trees were also made. Analysis of data and association rules were determined using Weka Tool.

4.1 Data Retrieval

While searching through various options in various databases such as DR-GAS, Uniprot, and other relevant databases, total 343 genes were obtained. Diseases data was also collected for both Endometrial Cancer and Colorectal Cancer from NCBI database.

4.2 Data Collection

Attributes of 215 genes were retrieved from the NCBI database using PHP program that include Mechanism, Mechanism details, Disease, Gene Type etc, then attributes of 129 genes were retrieved from the NCBI database and Mechanism details were also retrieved by searching it through web or from databases along with Gene Ontology information for all 343 genes using Gorilla tool.

Colorectal Cancer disease consists of 226 genes that includes the descriptions like Gene Id, description, start position of chromosome, end position of chromosome, chromosome no etc and 19 GO term were found using Gorilla software and for Endometrial Cancer consists of 85 genes and the same descriptions were collected in a excel sheet and it consists of 37 GO terms.

4.3 Decision Tree

Decision trees were made using the GO term and Gene name of all the data that was found and then the GO term falling on each node was computed and stored in a excel sheet and the same was done with when decision tree was made using gene name and gene names falling on each node. Analysis of data is performed using Rapid Miner and Weka Software to find what are the significant association rules and ranking of the attributes was performed and the same approach was applied on Diseases data.

5. Results and Discussions

a. The DNA Repair Genes and various information parameters

The DNA repair genes of Homo sapiens contains following information about 343 genes:

- **Repair gene Information** Which Includes Gene Id, Gene Name, Mechanism details, Gene type, other names, Description, Omim id, organism name, other designations, map location, chromosome , genomic nucleotide accession, version, start position on the genomic accession, end position on the genomic accession, orientation, and exon count.

- Decision tree of the data was made using the GO term and Gene name and stored that data in a excel sheet and this information was used for finding Association Rules.

This information will become a useful resource for DNA Repair Mechanism and can be used for research purpose. Figure 2 represents Decision tree of the DNA Repair data using the GO term , generated using the Rapid Miner software.

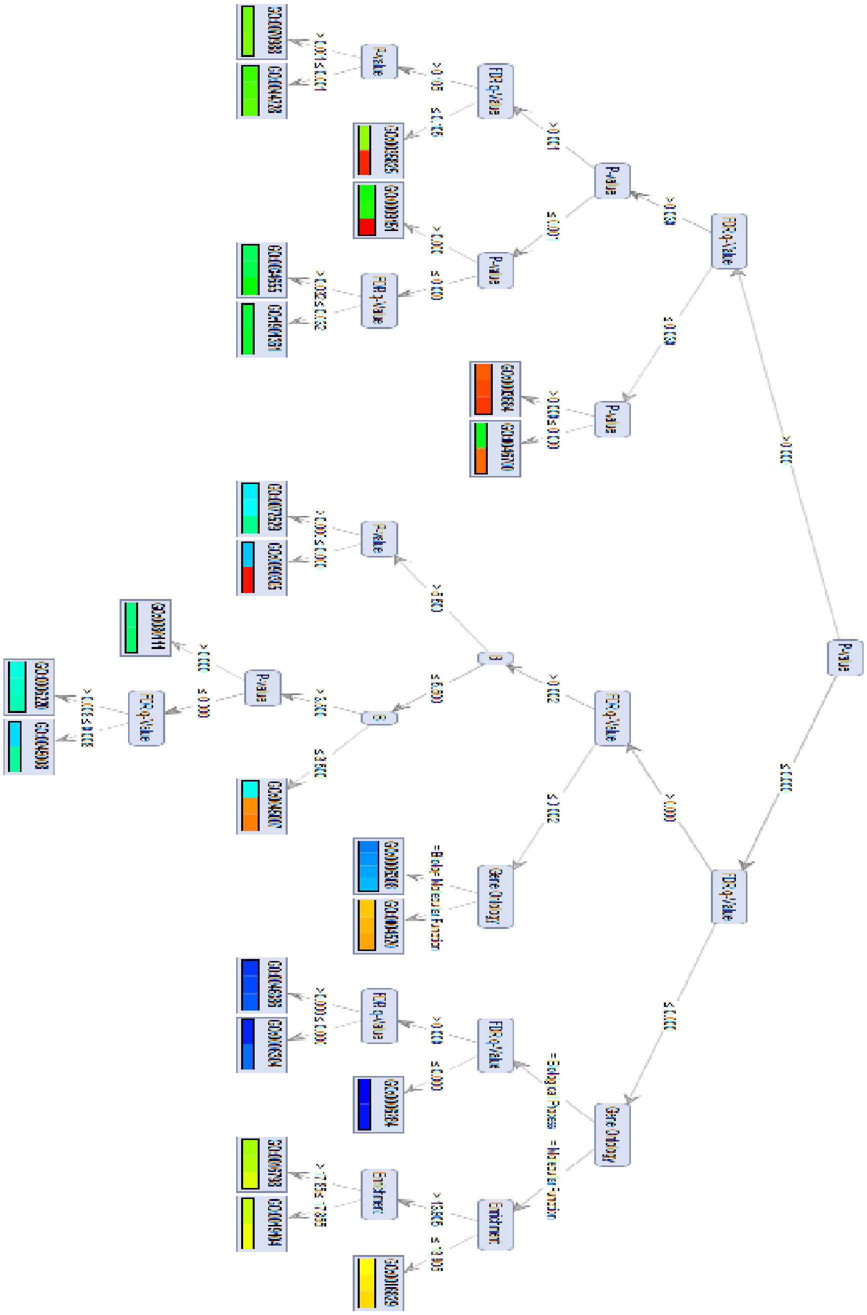

Fig 2: Decision Tree Using Go term

Similarly Figure 3 represests Decision tree of the DNA Repair data using the Gene name , generated using the Rapid Miner software.

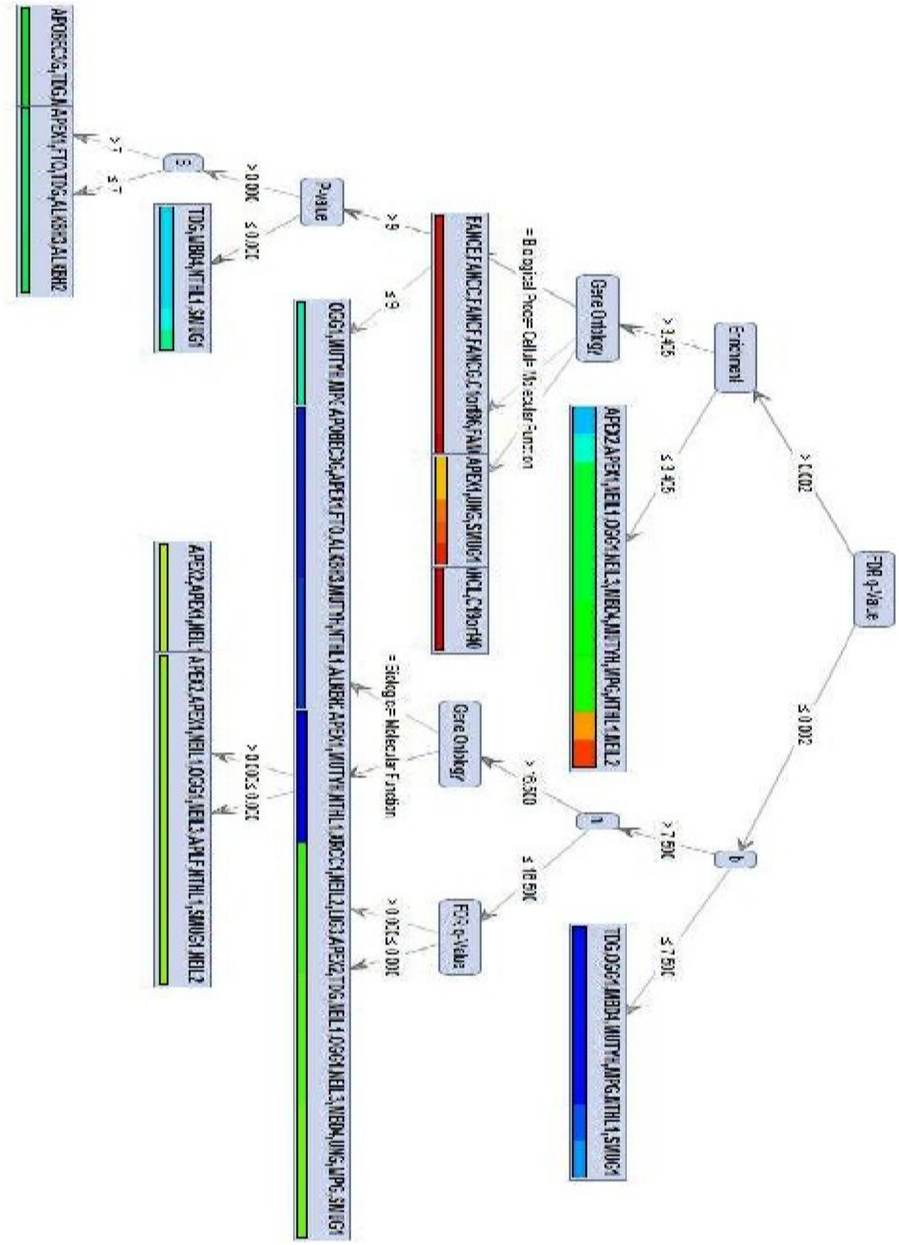

Fig 3: Decision Tree Using Gene Name

Decision Tree for the Disease: Colorectal Cancer

Figure 4 represents Decision tree of the Colorectal cancer data using the GO term , generated using the Rapid Miner software.

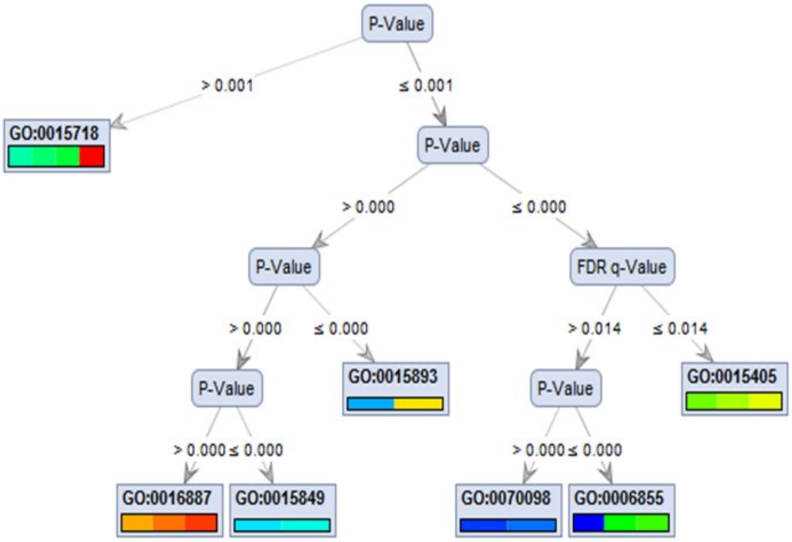

Fig 4:Decision Tree Using Go term

Disease: Endometrial Cancer Decision Tree

Figure 5 represents Decision tree of the endometrial cancer data using the GO term, generated using the Rapid Miner software.

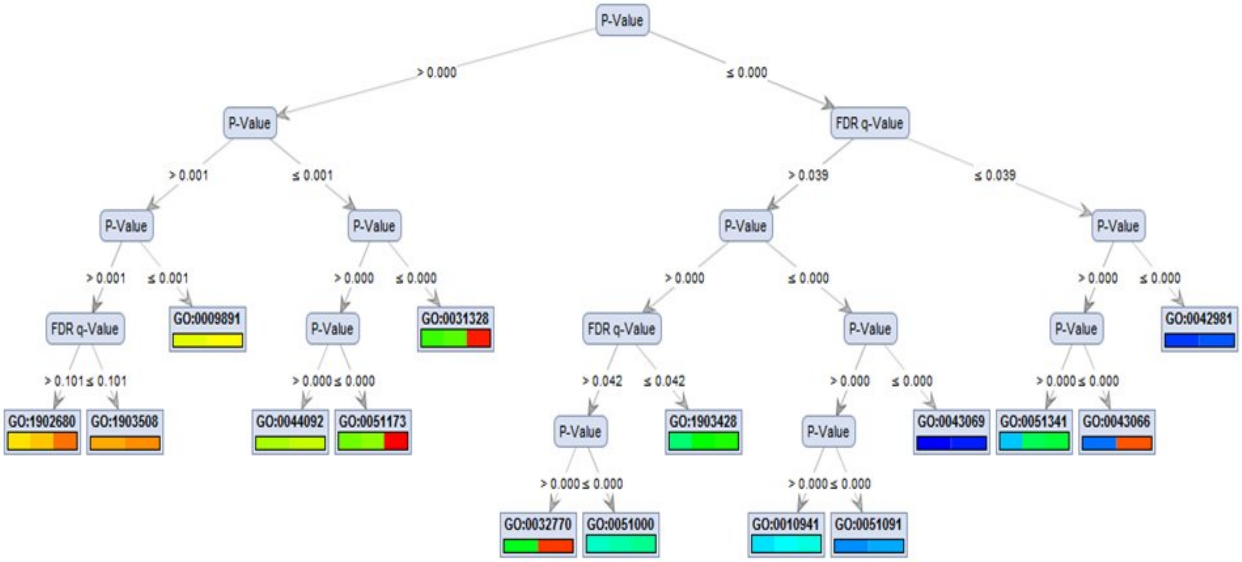

Fig 5: Decision Tree Using Go term

b. WEB-based GEne SeT AnaLysis Toolkit results:-

WebGestalt is a "WEB-based GEne SeT AnaLysis Toolkit". It is designed for functional genomics, proteomic and large-scale genetic studies from which large number of gene lists (e.g. differentially expressed gene sets, co-expressed gene sets etc) are continuously generated. WebGestalt incorporates information from different public resources and provides an easy way for biologists to make sense out of gene lists

- **Biological Process:-**

  X-axis=Biological Process

  Y-axis= Number of Genes

  Figure 6 represents Biological process of DNA repair genes data and maximum number of genes are associated with metabolic process.

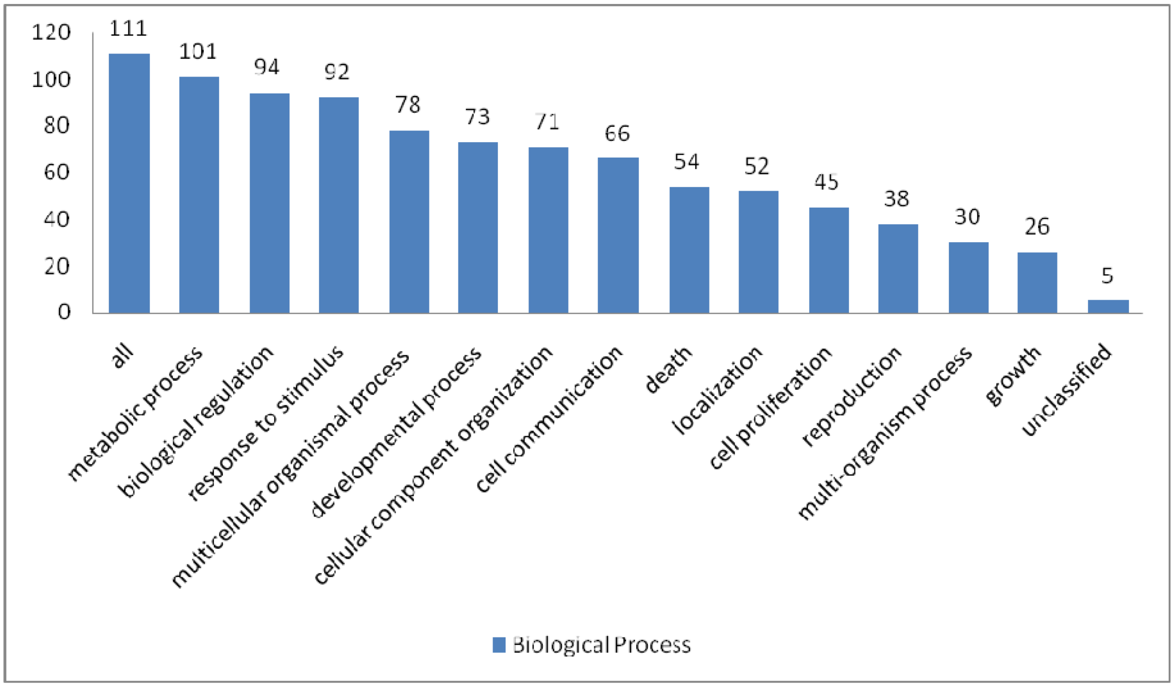

Fig 6: Biological process of DNA repair genes data

- **Molecular Function:-**

  X-axis=Molecular function

  Y-axis= Number of Genes

  Figure 7 represents Molecular Function of DNA repair genes data and maximum number of genes are associated with protein binding.

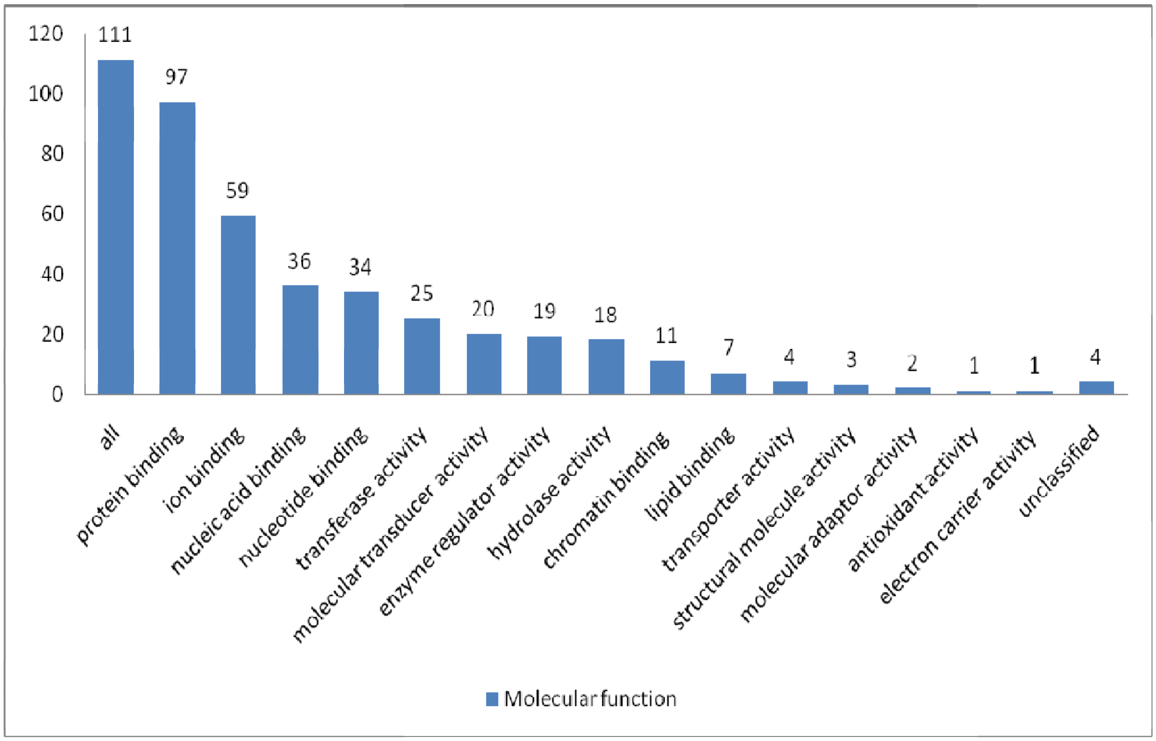

Fig 7: Molecular function of DNA repair genes data

- **Cellular Component:-**

  X-axis=Cellular Component

  Y-axis= Number of Genes

  Figure 8 represents the cellular component of DNA repair genes data and maximum number of genes are associated with nucleus.

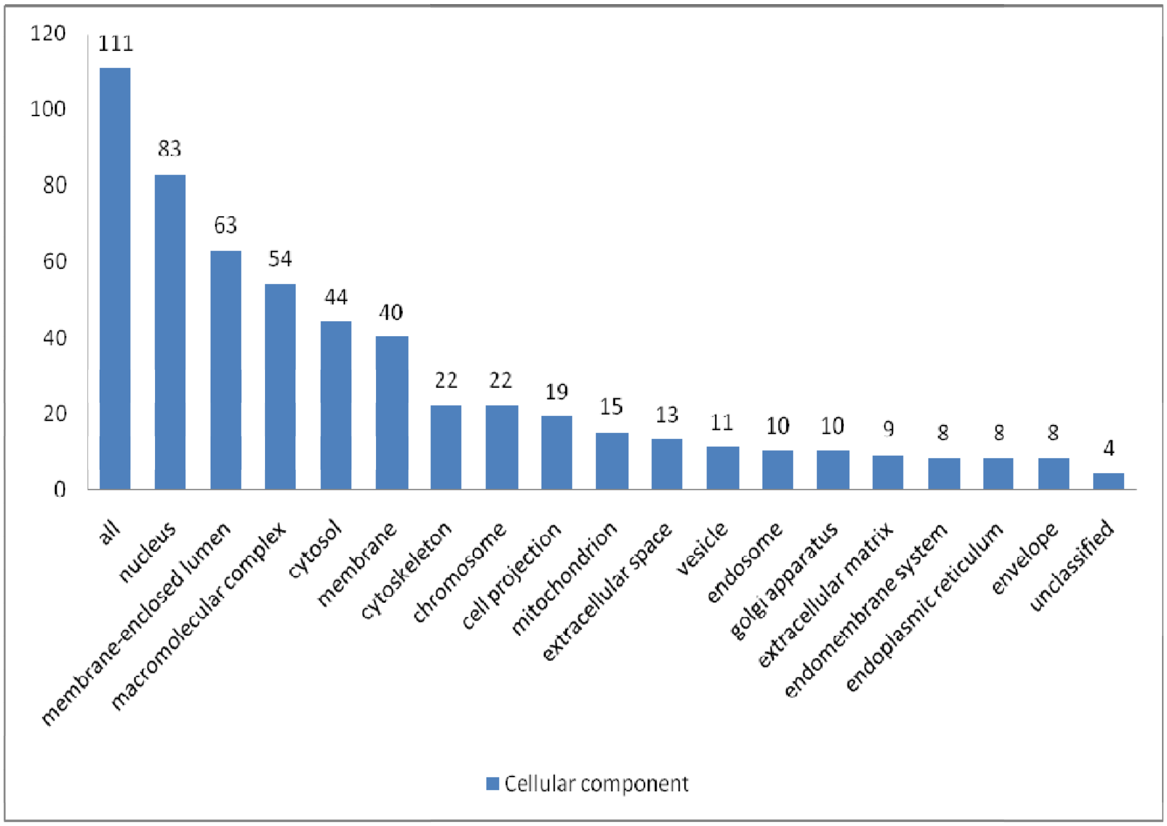

Fig 8: Cellular component of DNA repair genes data

Disease: Colorectal Cancer WebGestalt Results

Biological Process:-

X-axis=Biological Process

Y-axis= Number of Genes

Figure 9 represents Biological process of Colorectal cancer data and maximum number of genes are associated with response to stimulus.

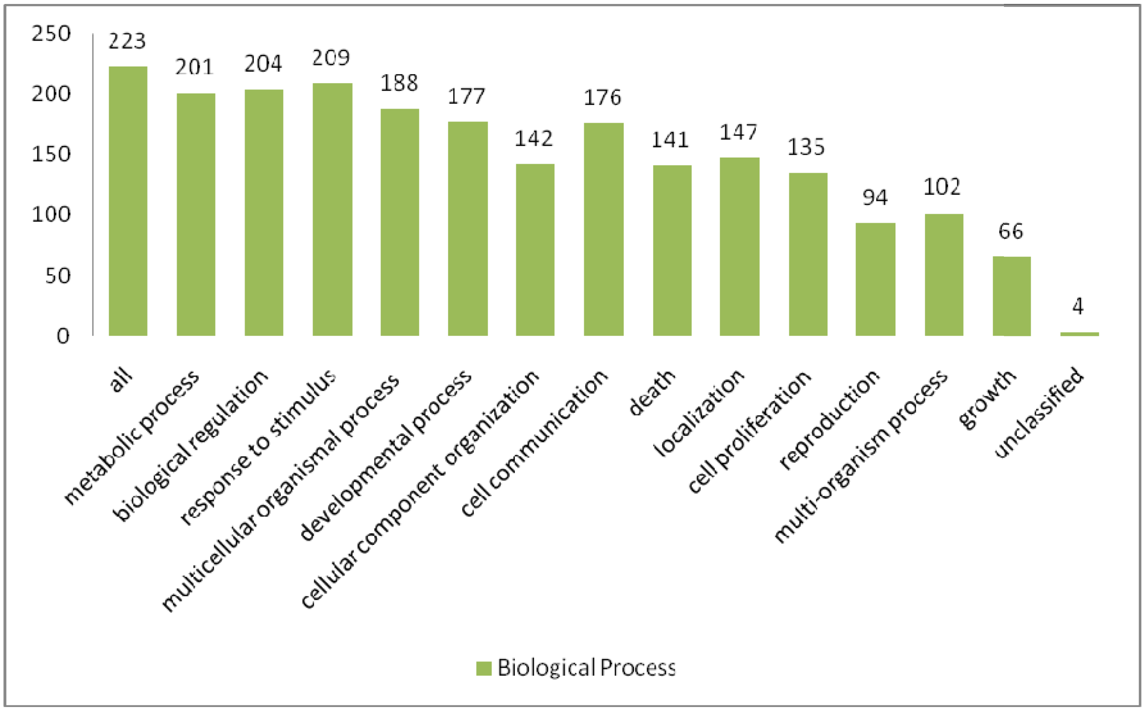

Fig 9:Biological process of Colorectal cancer data

Molecular Function:-

X-axis=Molecular function

Y-axis= Number of Genes

Figure 10 represents Molecular function of Colorectal cancer data and maximum number of genes are associated with protein binding.

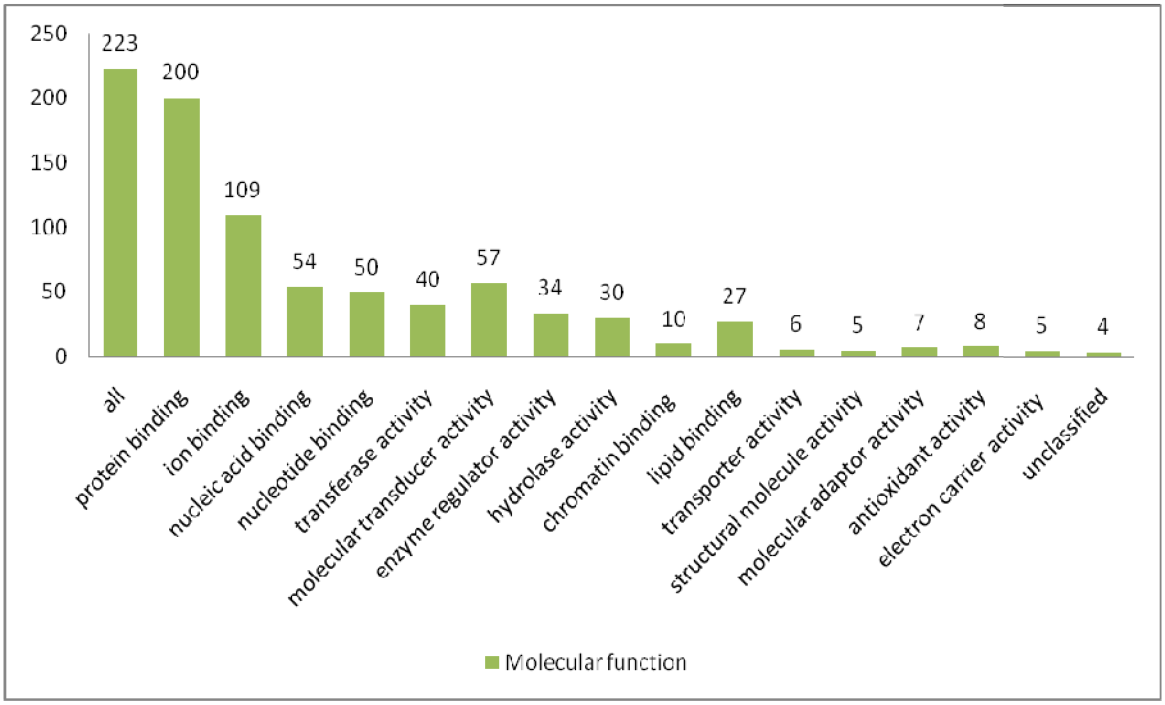

Fig 10:Molecular function of Colorectal cancer data

Cellular Component:-

X-axis=Cellular Component

Y-axis= Number of Genes

Figure 11 represents Cellular component of Colorectal cancer data and maximum number of genes are associated with membrane.

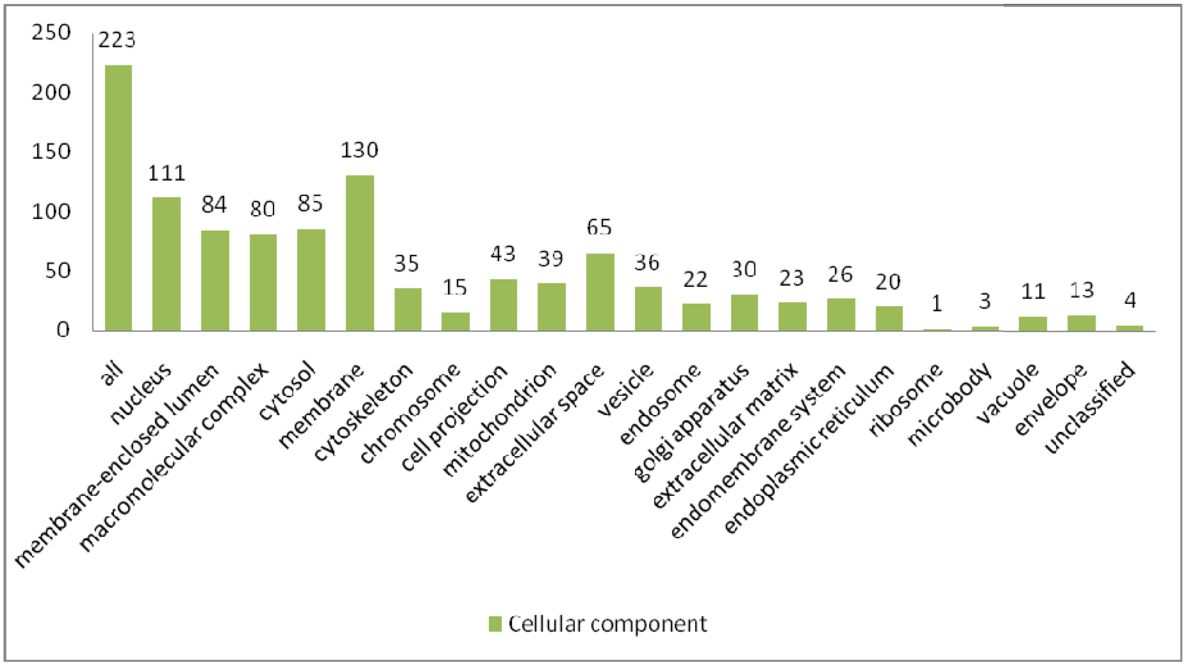

Fig 11: Cellular Component of Colorectal cancer data

Disease: Endometrial Cancer WebGestalt Results

Biological Process:-

X-axis=Biological Process

Y-axis= Number of Genes

Figure 12 represents Biological process of Endometrial cancer data and maximum number of genes are associated with biological regulation and response to stimulus.

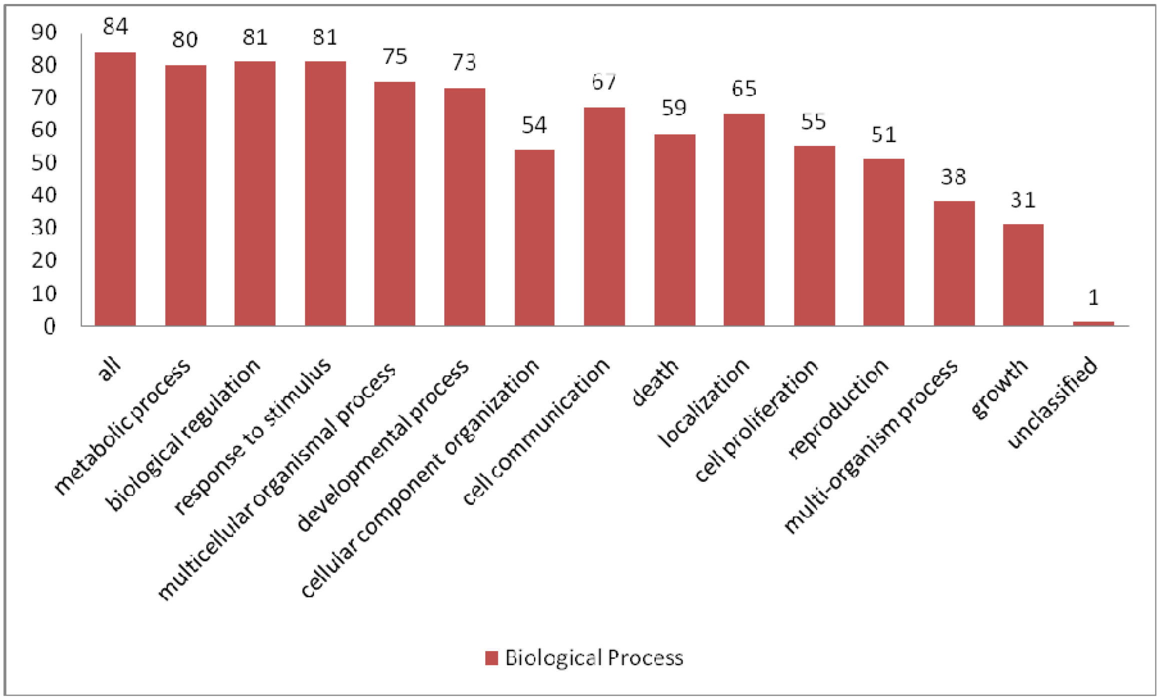

Fig 12: Biological process of Endometrial cancer data

Molecular Function:-

X-axis=Molecular function

Y-axis= Number of Genes

Figure 13 represents Molecular Function of Endometrial cancer data and maximum number of genes are associated with protein binding.

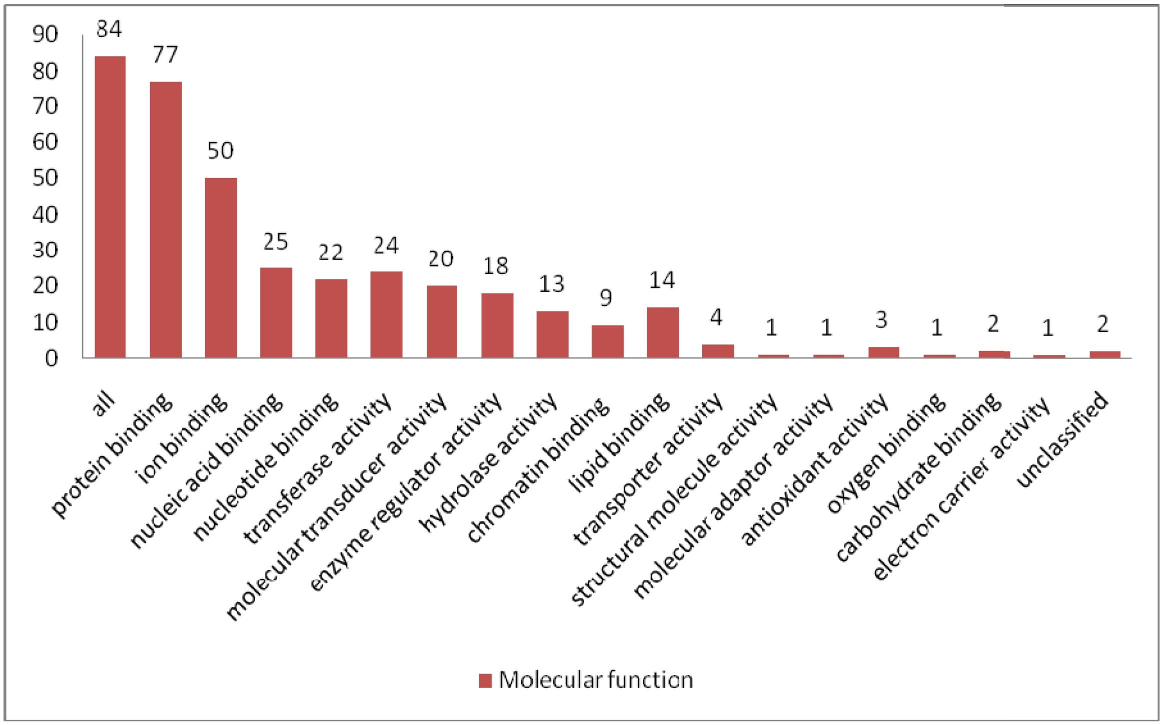

Fig 13: Molecular Function of Endometrial cancer data

Cellular Component:-

X-axis=Cellular Component

Y-axis= Number of Genes

Figure 14 represents Cellular component of Endometrial cancer data and maximum number of genes are associated with nucleus.

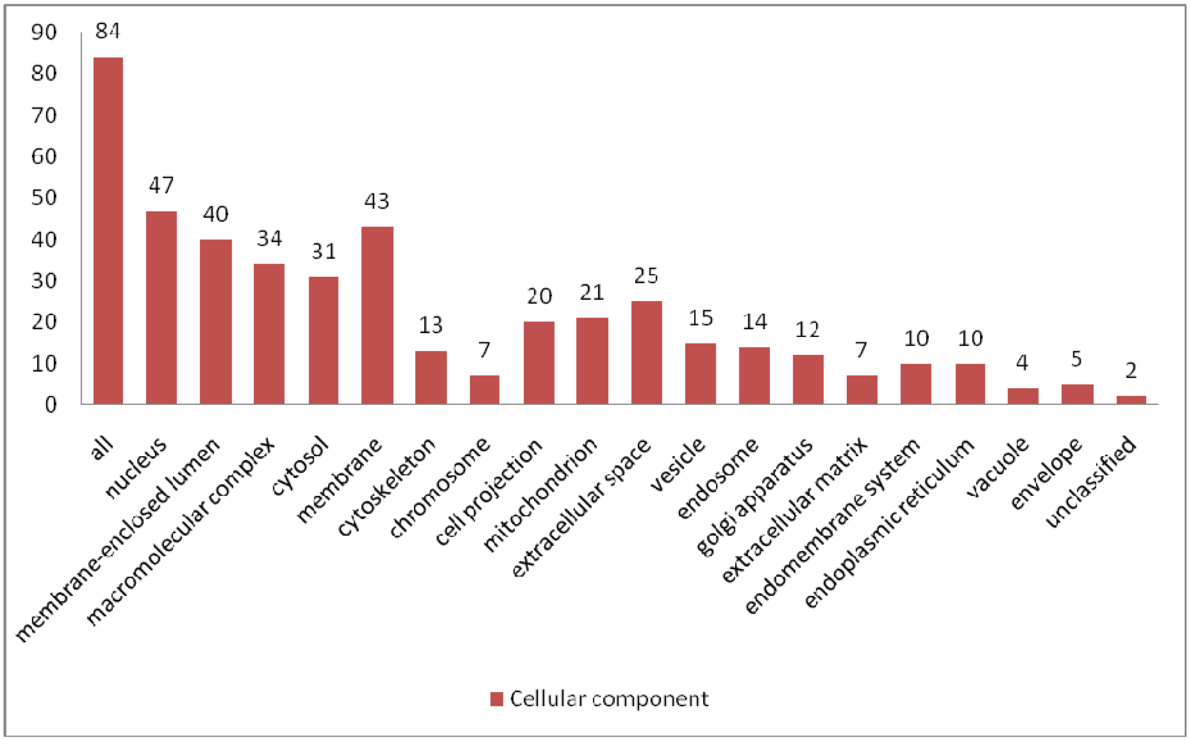

Fig 14: Cellular Component of Endometrial cancer data

Table 1:The summary table of decision Tree

Learning Algorithm	**C4.5 (Default)**
Attribute selection criterion	specifies the used method for selecting attribute, we choose gain ratio for this criterion
Minimal size for split	4
Minimal leaf size	1
Minimal gain	0.1
Maximal depth	20
Confidence value	0.25
Number of prepruning	3

c. Confusion Matrix

Table 2: Confusion Matrix

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.818	0.324	0.831	0.818	0.824	0.83	C0
	0.676	0.182	0.657	0.676	0.667	0.83	C1
Weighted Avg.	0.77	0.275	0.772	0.77	0.771	0.83	

d. Classifier Output

Table 3:Class for building and using a 0-R classifier.Predicts the mean (for a numeric class) or the mode (for a nominal class)

Decision Table	
NumberOfTrainingInstances	100
NumberOfRules	31
StartSet	noAttributes
SearchDirection	forward
Stale search after 5 node expansions	
TotalNumberOfSubsetsEvaluated	68
MeritOfBestSubsetFound	85
Evaluation (for feature Selection)	CV(Leave One Out)
Feature Set: 1,6,7,9,10,11	
CorrectlyClassifiedInstances	77%
IncorrectlyClassifiedInstances	23%
Kappa statistic	0.4912
Mean absolute error	0.3316
Root mean squared error	0.4002
Relative absolute error	73.6199 %
Root relative squared error	84.3668
Total Number of Instances	100

e. Ranking Attributes:-

Table 4:

GenerateRanking	True
numToSelect	-1
StartSet	
threshold	-1.7976931348623157E308

- **Chi-Squared Ranking**

Table 5:Evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class

Chi-Squared Ranking Filter	
Ranked Attributes	
9.216	Node7
2.564	Class1
1.375	Node3
1.327	Node6
0.932	Node8
0.557	Node2
0.391	Node5
0.204	Node9
0.188	Node4
Selected Attributes: 7,1,3,6,8,2,5,9,4:9	

- **Gain Ratio**

Table 6: Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.Gain Ratio (Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)

Gain Ratio Feature Evaluator	
Ranked Attributes	
0.07246	Node7
0.01858	Class1
0.00998	Node3
0.00976	Node6
0.00678	Node8
0.00409	Node2
0.00283	Node5
0.00149	Node9
0.00136	Node4
Selected Attributes: 7,1,3,6,8,2,5,9,4:9	

- **Exhaustive Search**

Table 7: Performs an exhaustive search through the space of attribute subsets starting from the empty set of attributes and reports the best subset found.

Exhaustive Search	
Start set	noAttributes
Number of evaluations	512
Merit of best subset found	0.07
Selected Attributes: 1(Class1),3(Node3),7(Node7),9(Node9) : 4	

f. Association Rule

Total of 10 rules are generated by using apriori algorithm for proper establishment of relations between different attributes. Minimum support threshold for this data is about 0.25 and confidence value is ≥90%.

Table 8: Class implementing an Apriori-type algorithm . Iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence. The algorithm has an option to mine class association rules.

car	False
classIndex	-1
delta	0.05
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.9
numRules	10
outputItemSets	False
removeAllMissingCols	False
significanceLevel	-1.0
upperBoundMinSupport	1.0
verbose	False

Association Rule Annotation

Out of 57 instances; Total of 10 rules are generated by using apriori algorithm for proper establishment of relations between different attributes. Minimum support threshold for this data is about 0.25 and confidence value is ≥90%.

All the 10 rules which are generated by taking different nodes; each implies one attributes give strong confidence for Gene Ontology. The nodes or attributes which belong to same class have high confidence value as compared to others.

'P-value' is the enrichment p-value computed according to the mHG or HG model.

'FDR q-value' is the correction of the above p-value for multiple testing using the Benjamini and Hochberg (1995) method.

Namely, for the ith term (ranked according to p-value) the FDR q-value is (p-value * number of GO terms) / i.

Enrichment (N, B, n, b) is defined as follows:

N - is the total number of genes

B - is the total number of genes associated with a specific GO term

n - is the number of genes in the top of the user's input list or in the target set when appropriate

b - is the number of genes in the intersection

Enrichment = (b/n) / (B/N)

I.	Description(Node3)^b(Node10)➔Gene Ontology(Node2)	conf:(0.97)
II.	Description(Node3)^n(Node9)➔ Gene Ontology(Node2)	conf:(0.97)
III.	P-value(Node4) ^b(Node10)➔N(Node7)	conf:(0.96)
IV.	Go Term(Class)^ Description(Node3)➔ Gene Ontology(Node2)	conf:(0.96)
V.	Description(Node3)^B(Node8)➔Gene Ontology(Node2)	conf:(0.94)

VI. Gene Ontology(Node2)^ b(Node10)➔ Description(Node3) conf:(0.94)
VII. Description(Node3)^FDR q-value(Node5)➔ Gene Ontology(Node2) conf:(0.94)
VIII. Gene Ontology(Node2)^ B(Node8)➔ Description(Node3) conf:(0.94)
IX. Description(Node3)^P-value(Node4)➔Gene Ontology(Node2) conf:(0.93)
X. Description(Node3)^Enrichment(Node6)➔ B(Node8) conf:(0.93)

Disease: Colorectal Cancer Weka Results

Confusion Matrix

Table 9**:** Confusion Matrix

	TP Rate	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**	**Class**
	0.788	0.25	0.774	0.788	0.781	0.839	Value1
	0.75	0.212	0.766	0.75	0.758	0.839	Value2
Weighted Avg.	0.77	0.232	0.77	0.77	0.77	0.839	

Classifier Output

Table 10: Class for building and using a 0-R classifier. Predicts the mean (for a numeric class) or the mode (for a nominal class)

Decision Table	
NumberOfTrainingInstances	100
NumberOfRules	16
StartSet	noAttributes
SearchDirection	forward
Stale search after 5 node expansions	
TotalNumberOfSubsetsEvaluated	76
MeritOfBestSubsetFound	75
Evaluation (for feature Selection)	CV(Leave One Out)
Feature Set: 4,5,7,8,10	
CorrectlyClassifiedInstances	77%
IncorrectlyClassifiedInstances	23%
Kappa statistic	0.5389
Mean absolute error	0.3491
Root mean squared error	0.4015
Relative absolute error	69.9229 %
Root relative squared error	80.3561%
Total Number of Instances	100

Ranking Attributes:-

- ### Chi-Squared Ranking

Table 11: Evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class

Chi-Squared Ranking Filter	
Ranked Attributes	
9.216	Node7
2.564	Class1
1.375	Node3
1.327	Node6
0.932	Node8
0.557	Node2
0.391	Node5
0.204	Node9
0.188	Node4
Selected Attributes: 7,1,3,6,8,2,5,9,4:9	

- ### Gain Ratio

Table 12: Evaluates the worth of an attribute by measuring the gain ratio with respect to the class. Gain Ratio (Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)

Gain Ratio Feature Evaluator	
Ranked Attributes	
0.07246	Node7
0.01858	Class1
0.00998	Node3
0.00976	Node6
0.00678	Node8
0.00409	Node2
0.00283	Node5
0.00149	Node9
0.00136	Node4
Selected Attributes: 7,1,3,6,8,2,5,9,4:9	

- **Exhaustive Search**

Table 13: Performs an exhaustive search through the space of attribute subsets starting from the empty set of attributes and reports the best subset found.

Exhaustive Search	
Start set	noAttributes
Number of evaluations	512
Merit of best subset found	0.07
Selected Attributes: 1(Class1),3(Node3),7(Node7),9(Node9) : 4	

Association Rule

Class implementing an Apriori-type algorithm . Iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence. The algorithm has an option to mine class association rules.

I.	Description(Node3)^b(Node10)➔Gene Ontology(Node2)	conf:(0.97)
II.	Description(Node3)^n(Node9)➔ Gene Ontology(Node2)	conf:(0.97)
III.	P-value(Node4) ^b(Node10)➔N(Node7)	conf:(0.96)
IV.	Go Term(Class)^ Description(Node3)➔ Gene Ontology(Node2)	conf:(0.96)
V.	Description(Node3)^B(Node8)➔Gene Ontology(Node2)	conf:(0.94)
VI.	Gene Ontology(Node2)^ b(Node10)➔ Description(Node3)	conf:(0.94)
VII.	Description(Node3)^FDR q-value(Node5)➔ Gene Ontology(Node2)	conf:(0.94)
VIII.	Gene Ontology(Node2)^ B(Node8)➔ Description(Node3)	conf:(0.94)
IX.	Description(Node3)^P-value(Node4)➔Gene Ontology(Node2)	conf:(0.93)
X.	Description(Node3)^Enrichment(Node6)➔ B(Node8)	conf:(0.93)

Disease: Endometrial Cancer Weka Results

Confusion Matrix

Table 14:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.788	0.25	0.774	0.788	0.781	0.839	Value1
	0.75	0.212	0.766	0.75	0.758	0.839	Value2
Weighted Avg.	0.77	0.232	0.77	0.77	0.77	0.839	

Classifier Output

Table 15: Class for building and using a 0-R classifier. Predicts the mean (for a numeric class) or the mode (for a nominal class)

Decision Table	
NumberOfTrainingInstances	100
NumberOfRules	16
StartSet	noAttributes
SearchDirection	forward
Stale search after 5 node expansions	
TotalNumberOfSubsetsEvaluated	76
MeritOfBestSubsetFound	75
Evaluation (for feature Selection)	CV(Leave One Out)
Feature Set: 4,5,7,8,10	
CorrectlyClassifiedInstances	77%
IncorrectlyClassifiedInstances	23%
Kappa statistic	0.5389
Mean absolute error	0.3491
Root mean squared error	0.4015
Relative absolute error	69.9229 %
Root relative squared error	80.3561%
Total Number of Instances	100

Ranking Attributes:-

- ### Chi-Squared Ranking

Table 16: Evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class

Chi-Squared Ranking Filter	
Ranked Attributes	
9.216	Node7
2.564	Class1
1.375	Node3
1.327	Node6
0.932	Node8
0.557	Node2
0.391	Node5
0.204	Node9
0.188	Node4
Selected Attributes: 7,1,3,6,8,2,5,9,4:9	

- **Gain Ratio**

Table 17: Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.
Gain Ratio (Class, Attribute) = (H(Class) - H(Class | Attribute)) / H(Attribute)

Gain Ratio Feature Evaluator	
Ranked Attributes	
0.07246	Node7
0.01858	Class1
0.00998	Node3
0.00976	Node6
0.00678	Node8
0.00409	Node2
0.00283	Node5
0.00149	Node9
0.00136	Node4
Selected Attributes: 7,1,3,6,8,2,5,9,4:9	

- **Exhaustive Search**

Table 18: Performs an exhaustive search through the space of attribute subsets starting from the empty set of attributes and reports the best subset found.

Exhaustive Search	
Start set	noAttributes
Number of evaluations	512
Merit of best subset found	0.07
Selected Attributes: 1(Class1),3(Node3),7(Node7),9(Node9) : 4	

Association Rule

Class implementing an Apriori-type algorithm . Iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence. The algorithm has an option to mine class association rules.

I. Description(Node3)^b(Node10)➔Gene Ontology(Node2) conf:(0.97)
II. Description(Node3)^n(Node9)➔ Gene Ontology(Node2) conf:(0.97)
III. P-value(Node4) ^b(Node10)➔N(Node7) conf:(0.96)
IV. Go Term(Class)^ Description(Node3)➔ Gene Ontology(Node2) conf:(0.96)
V. Description(Node3)^B(Node8)➔Gene Ontology(Node2) conf:(0.94)
VI. Gene Ontology(Node2)^ b(Node10)➔ Description(Node3) conf:(0.94)
VII. Description(Node3)^FDR q-value(Node5)➔ Gene Ontology(Node2) conf:(0.94)

VIII. Gene Ontology(Node2)^ B(Node8)➔ Description(Node3) conf:(0.94)

IX. Description(Node3)^P-value(Node4)➔Gene Ontology(Node2) conf:(0.93)

X. Description(Node3)^Enrichment(Node6)➔ B(Node8) conf:(0.93)

Conclusion and Future Prospects

It was found that mostly the DNA repair genes are associated with metabolic process and performing the protein binding function and associated with nucleus. With the help of Association rule we come to know

- Out of all association rules that we have got, majority of association rules were found with Description→Gene Ontology with high confidence value of 0.97

From this it is concluded is that for a particular type of description and Gene ontology what are the possible set of genes that we are going to have.

Similarly we applied same approach to selected diseases colorectal cancer and endometrial cancer and we have selected these two diseases only because out of all Mechanism that we have, we randomly selected any two mechanism mainly BER and MMR mechanism and these two diseases were found to be prominent in this and the results are:-

- In case of Colorectal cancer, out of all association rule that we have got majority of association rules were found with Description→Gene Ontology and from this it is concluded is that for a particular type of description and Gene ontology what are the possible set of genes that we are going to have.

- In case of Endometrial Cancer, out of all association rule that we have got majority of association rules were found with Description→Gene Ontology and from this it is concluded is that for a particular type of description and Gene ontology what are the possible set of genes that we are going to have.

These two diseases results were also following the same results as that we have got under DNA repair genes. It means that we can determine the possible set of genes if we know the Gene ontology term and Description of any data.

6. References

a. Journals References

1. Sehgal M.,and Singh T.R., DR-GAS: A database of functional genetic variants and their phosphorylation states in human DNA repair systems, DNA Repair, 16: 97-103

2. Bjorksten, J; Acharya, PV; Ashman, S; Wetlaufer, DB (1971). "Gerogenic fractions in the tritiated rat.". Journal of the American Geriatrics Society 19 (7): 561–74

3. Browner, WS; Kahn, AJ; Ziv, E; Reiner, AP; Oshima, J; Cawthon, RM; Hsueh, WC; Cummings, SR. (2004). "The genetics of human longevity". Am J Med 117 (11): 851–60.

4. Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106.

5. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

6. World Cancer Report 2014. World Health Organization. 2014. pp. Chapter 5.5. ISBN 9283204298.

7. W. Peng, J. Chen, and H. Zhou, “An Implementation of ID3 - Decision Tree Learning Algorithm,” From web. arch. usyd. edu. au/wpeng/ …. Sydney, Australia, 2009.

8. "SEER Stat Fact Sheets: Endometrial Cancer". National Cancer Institute. Retrieved 18 June 2014.

9. World Cancer Report 2014. World Health Organization. 2014. pp. Chapter 5.5.

10. Soliman, PT; Lu, KH (2013). "Neoplastic Diseases of the Uterus". In Lentz, GM; Lobo, RA; Gershenson, DM; Katz, VL. Comprehensive Gynecology (6th ed.).

11. I. Jenhani, N. B. Amor, and Z. Elouedi, “Decision Trees as Possibilistic Classifiers,” International Journal of Approximate Reasoning, vol. 48, no. 3, pp. 784–807, Aug. 2008.

12. A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection,” International Journal of Information Technology and Knowledge Management, vol. 2, no. 2, pp. 271–277, 2010.

13. J. Han and M. Kamber, Data Mining Concepts and Techniques, Second Edi. 2006, p. 743.

14. R. Kohavi and R. Quinlan, “Decision Tree Discovery,” Citeseer, vol. 3, 1999.

15. "Endometrial Cancer Prevention". PDQ. NIH. 28 February 2014.

16. Hoeijmakers JH (2009) DNA damage, aging, and cancer. N Engl J Med 361(15):1475–1485

17. Stracker TH, Usui T, Petrini JH (2009) Taking the time to make important decisions: the checkpoint effector kinases Chk1 and Chk2 and the DNA damage response. DNA Repair (Amst) 8(9):1047–1054

18. Zhou BB, Elledge SJ (2000) The DNA damage response: putting checkpoints in perspective. Nature 408(6811):433–439

19. Rich T, Allen RL,WyllieAH(2000) Defying death afterDNAdamage. Nature 407(6805):777–783.

20. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362(6422):709–715

21. LindahlT, NybergB(1972) Rate of depurination of native deoxyribonucleic acid. Biochemistry 11(19):3610–3618

22. Sugiyama H, Fujiwara T, Ura A et al (1994) Chemistry of thermal degradation of abasic sites in DNA. Mechanistic investigation on thermal DNA strand cleavage of alkylated DNA. ChemRes Toxicol 7(5):673–683

23. Weerakkody, Yuranga; Gaillard, Frank. "Colorectal carcinoma". Radiopaedia.org. Retrieved 13 September 2014

24. “What You Need To Know: Endometrial Cancer”. NCI. National Cancer Institute. Retrieved 6 August 2014

25. Krokan HE, Drablos F, Slupphaug G (2002) Uracil in DNA–occurrence, consequences and repair. Oncogene 21(58):8935–8948

26. KowYW (2002) Repair of deaminated bases in DNA. Free Radic Biol Med 33(7):886–893

27. Apel K, Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annu Rev Plant Biol 55:373–399

28. Marnett LJ (2000) Oxyradicals and DNA damage. Carcinogenesis 21(3):361–370

29. Cadet J, Berger M, Douki T, Ravanat JL (1997) Oxidative damage to DNA: formation, measurement, and biological significance. Rev Physiol Biochem Pharmacol 131:1–87

30. Burney S, Caulfield JL, Niles JC,Wishnok JS, Tannenbaum SR (1999) The chemistry of DNA damage from nitric oxide and peroxynitrite. Mutat Res 424(1–2):37–49

31. Ravanat J-L (2005) Measuring oxidized DNA lesions as biomarkers of oxidative stress: an analytical challenge FABAD. J Pharm Sci 30(2):100–113

32. Major GN, Collier JD (1998) Repair of DNA lesion O6-methylguanine in hepatocellular carcinogenesis. J Hepatobiliary Pancreat Surg 5(4):355–366

33. McCulloch SD, Kunkel TA (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. Cell Res 18(1):148–161

34. Shimizu M, Gruz P, KamiyaHet al (2003) Erroneous incorporation of oxidizedDNAprecursors byY-family DNA polymerases. EMBO Rep 4(3):269–273

35. Fortini P, Dogliotti E (2007) Base damage and single-strand break repair: mechanisms and functional significance of short- and long-patch repair subpathways. DNA Repair (Amst) 6(4):398–409

36. CaldecottKW(2003) XRCC1 andDNAstrand break repair. DNARepair (Amst) 2(9):955–969

37. Marintchev A, Mullen MA, Maciejewski MW, Pan B, Gryk MR, Mullen GP (1999) Solution structure of the single-strand break repair protein XRCC1 N-terminal domain. Nat Struct Biol6(9):884–893

38. Malanga M, Althaus FR (2005) The role of poly(ADP-ribose) in the DNA damage signaling network. Biochem Cell Biol 83(3):354–364

39. Peltomaki P (2001) Deficient DNA mismatch repair: a common etiologic factor for colon cancer. Hum Mol Genet 10(7):735–740

40. Li GM (2008) Mechanisms and functions of DNA mismatch repair. Cell Res 18(1):85–98

41. Fukui K (2010) DNA mismatch repair in eukaryotes and bacteria. J Nucleic Acids 2010:1–6

42. Larrea AA, Lujan SA, Kunkel TA (2010) SnapShot: DNA mismatch repair. Cell 141(4):730 e1

43. Modrich P (2006) Mechanisms in eukaryotic mismatch repair. J Biol Chem 281(41):30305–30309

44. Galio L, Bouquet C, Brooks P (1999)ATP hydrolysis-dependent formation of a dynamic ternary nucleoprotein complex with MutS and MutL. Nucleic Acids Res 27(11):2325–2331

45. Tran PT, Erdeniz N, Symington LS, Liskay RM (2004) EXO1-A multi-tasking eukaryotic nuclease. DNA Repair (Amst) 3(12):1549–1559

46. Kadyrov FA, Holmes SF, Arana ME et al (2007) Saccharomyces cerevisiae MutLalpha is a mismatch repair endonuclease. J Biol Chem 282(51):37181–37190

47. Shuck SC, Short EA, Turchi JJ (2008) Eukaryotic nucleotide excision repair: from understanding mechanisms to influencing biology. Cell Res 18(1):64–72

48. Costa RM, Chigancas V, Galhardo Rda S, Carvalho H, Menck CF (2003) The eukaryotic nucleotide excision repair pathway. Biochimie 85(11):1083–1099

49. Nouspikel T (2008) Nucleotide excision repair and neurological diseases. DNA Repair (Amst) 7(7):1155–1167

b. Web References

W1. NCBI
http://www.ncbi.nlm.nih.gov/
W2. DR-GAS Database
http://www.bioinfoindia.org/drgas/
W3. WEB-based GEne SeT AnaLysis Toolkit
http://bioinfo.vanderbilt.edu/webgestalt/
W4. Gorilla Tool
http://cbl-gorilla.cs.technion.ac.il/
W5. Repairtoire- A database of DNA repair pathways
http://repairtoire.genesilico.pl/