DOKUZ EYLÜL UNIVERSITY
ENGINEERING FACULTY
DEPARTMENT OF COMPUTER ENGINEERING

PREDICTION OF NEW TREATMENTS FOR PANCREATIC CANCER

by
Can KIZILÖZ
Ece KOBANÇ

Advisor
Assoc. Prof. Dr. Zerrin IŞIK

June, 2022
İZMİR

PREDICTION OF NEW TREATMENTS FOR PANCREATIC CANCER

A Thesis Submitted to the
Dokuz Eylül University, Department of Computer Engineering
In Partial Fulfillment of the Requirements for the Degree of B.Sc.

by

Can KIZILÖZ
Ece KOBANÇ

Advisor
Assoc. Prof. Dr. Zerrin IŞIK

June, 2022
İZMİR

SENIOR PROJECT EXAMINATION RESULT FORM

We have read the thesis entitled **"PREDICTIONS OF NEW TREATMENTS FOR PANCREATIC CANCER "** completed by **Can KIZILÖZ** and **Ece KOBANÇ** under advisor of **Assist. Prof. Dr. Zerrin IŞIK** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of B.Sc.

_____

_____

Advisor

_____ _____

_____ _____

Committee Member Committee Member

_____

Prof. Dr. Yalçın ÇEBİ

Chair

Department of Computer Engineering

ACKNOWLEDGEMENTS

We would like to thank our consultant Assoc. Prof. Dr. Zerrin IŞIK for his continuous assistance at all stages of our project. We are deeply grateful for her guidance, support, patience and motivation during difficult times. This project would not have been completed without her.

Can KIZILÖZ
Ece KOBANÇ

PREDICTION OF NEW TREATMENTS FOR PANCREATIC CANCER

ABSTRACT

Pancreatic cancer has a rather poor prognosis, with a five-year survival rate of less than 5% of patients diagnosed. The most important reasons for this prognosis are that approximately 90% of patients are diagnosed at stage III or IV as pancreatic cancer progresses with almost no clinical symptoms and the therapeutic options available are limited. The aim of the project is to find candidate treatments for pancreas with appropriate pre-processing methods and algorithms using microarray gene expression profiles. The project also offers a platform where all users, who have gene expression data in cancer-specific Affymetrix U133 Plus 2.0 whole-genome chip, can search for candidate drugs using their own data. Within the scope of project, the most convenient pancreas cancer samples and healthy controls were searched, two of the data sets are selected. Differential expression analysis was performed on both data sets by applying statistical tests. We obtained functional and physical protein interactions from the STRING database to create a protein-protein interaction (PPI) network. In order to identify the most important genes in this network, the personalized Page Rank algorithm was run on the network. The most important genes were chosen from the high rank valued ones, then they were matched with differentially expressed genes obtained from the microarray data analysis. The rDGIdb library was used for searching of drugs that are targeting treatment candidates. As a result of this search, a total of 71 drugs were obtained as new suggestions for pancreas cancer treatment.

After receiving the results of our study, the above-mentioned algorithms were rearranged to work on microarray data set. Using NodeJs, HTML, CSS, JavaScript web technologies, a web environment was developed where users can upload their cancer-specific datasets as normal samples and tumor samples. In addition, the user can enter his own p and fold change values here.

As a conclusion, this project both provides new candidate treatments for pancreas cancer and web platform for users to analyze their own microarray data for finding potential treatments for different diseases.

PANKREAS KANSERİNDE YENİ TEDAVİLERİN TAHMİNİ

ÖZET

Pankreas kanseri, teşhis konulan hastaların %5'inden daha az beş yıllık sağkalım oranı ile oldukça kötü bir prognoza sahiptir. Bu prognozun en önemli nedenleri, pankreas kanserinin neredeyse hiç klinik semptom göstermeden ilerlemesi ve mevcut tedavi seçeneklerinin sınırlı olması nedeniyle hastaların yaklaşık %90'ının evre III veya IV'te teşhis edilmesidir. Projenin amacı, mikrodizi gen ekspresyon profillerini kullanarak uygun ön işleme yöntemleri ve algoritmaları ile pankreas için aday tedavileri bulmaktır. Proje ayrıca kansere özgü Affymetrix U133 Plus 2.0 tam genom çipinde gen ekspresyonu verilerine sahip tüm kullanıcıların kendi verilerini kullanarak aday ilaçları arayabileceği bir platform sunuyor. Proje kapsamında en uygun pankreas kanseri örnekleri ve sağlıklı kontroller aranmış, veri setlerinden iki tanesi seçilmiştir. Her iki veri seti üzerinde istatistiksel testler uygulanarak diferansiyel ifade analizi yapılmıştır. Bir protein-protein etkileşimi (PPE) ağı oluşturmak için STRING veri tabanından fonksiyonel ve fiziksel protein etkileşimlerini elde ettik. Bu ağdaki en önemli genleri belirlemek için ağ üzerinde kişiselleştirilmiş Page Rank algoritması çalıştırıldı. En önemli genler, yüksek sıra değerli olanlardan seçildi, daha sonra mikrodizi veri analizinden elde edilen diferansiyel olarak eksprese edilen genlerle eşleştirildi. Tedavi adaylarını hedefleyen ilaçların aranması için rDGIdb kütüphanesi kullanıldı. Bu araştırma sonucunda pankreas kanseri tedavisi için yeni öneriler olarak toplam 71 ilaç elde edildi. Çalışmamızın sonuçlarını aldıktan sonra, yukarıda belirtilen algoritmalar her veri seti üzerinde çalışacak şekilde yeniden düzenlendi.

NodeJs, HTML, CSS, JavaScript web teknolojileri kullanılarak, kullanıcıların mikrodizi veri örneklerini yükleyebilecekleri bir web ortamı geliştirildi. Ayrıca kullanıcı kendi p ve t değerlerini buraya girebilir.

Sonuç olarak, bu proje hem pankreas kanseri için yeni aday tedaviler hem de kullanıcıların farklı hastalıklar için potansiyel tedaviler bulmak için kendi mikrodizi verilerini analiz etmeleri için bir web platformu sunmaktadır.

CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

1.1 Background Information

Pancreatic cancer maintains as one of the most dangerous cancer types, with the highest death rate among all major cancers. Currently, gemcitabine is the only therapeutic drug approved for treating pancreatic cancer, yet its response rate is poor. Hence, effective and robust treatments are urgently needed for treating pancreatic cancer. Drug development has never been easy as it is a time-consuming process and any failure leads to the returning of the beginning of the clinical trial phase which is also expensive. Most of the studies cannot even make to proceed it to the phase trial processes, unfortunately. Prediction of a new drug with machine learning algorithms comes as the most efficient solution for these situations. Consumed time and budget can be reduced by eliminating extra trials because of the perfect fitting, effective and robust solution which is found by the prediction algorithms.

Diversified approaches are being used to obtain the therapeutic drugs, but the most efficient way is the computational ones which includes machine learning algorithms. The most common way for predicting therapeutic drugs is using the combination of different techniques in computing, e.g., searching of a negative correlation between drug induced gene expression profiles and differentially expressed genes observed in a target disease. We think that our approaches on this method will uncover the best therapeutic drug which is effective and promising alternative for treating pancreatic cancer.

1.2 Problem Definition

Pancreatic cancer has a rather poor prognosis with a five-year survival rate of less than 5% of approximately diagnosed patients. The reasons for this prognosis include that approximately 90% of patients are diagnosed at stage III or IV, as pancreatic cancer progresses almost without any clinical symptoms, and the available therapeutic options are limited . (Kikuyama et al, 2018)

Development of a new drug covers understanding a disease processes thoroughly, evaluating the ideas for curing the disease, screening a large number of compounds to find possible beneficial effects, genetic and protein tests of the screened compounds, safety tests, FDA approval, and monitoring of the drug after approval phases. The entire development pipeline involves several challenging and very complex processes. The process of developing a new drug is costly and time-consuming. The most effective way to shorten this process is repositioning of approved drugs for treating new diseases. Drug repositioning significantly reduces R&D costs. The present compounds have demonstrated safety in humans and do not require Phase I trials. Thus, it can be marketed more quickly and cost-effectively. Drug repositioning approaches are often based on gene expression. If two drugs exhibit similar gene expression patterns, it can be interpreted that they have similar therapeutic effects.

This project will look for an answer this hypothesis: If gene expression patterns of several drugs and pancreas cancer-specific gene expression patterns are obtained from databases, would a negative correlation between two expression patterns suggest a new candidate for cancer treatment? Additional analyses are needed for this approach to be meaningful.

1.3 Motivation / Related Works

According to the GLOBOCAN 2020 cancer incidence and mortality estimates produced by the International Agency for Research on Cancer, 466.000 people died out of 496.000 diagnosed cases of pancreatic cancer in 2020, and pancreatic cancer is the seventh leading cause of cancer death in both sexes. (Sung et al, 2021) According to the American Cancer Society's estimates of pancreatic cancer in the United States for 2021, approximately 60.430 people

(31.950 men and 28.480 women) will be diagnosed with pancreatic cancer and approximately 48.220 people (25.270 men and 22.950 women) will die of pancreatic cancer in 2021. (American Cancer Society, 2021) Pancreatic cancer is more common in men than in women. Globally, the incidence of pancreatic cancer is 5,5 per 100.000 for men and 4,0 per 100.000 for women. (Rawla et al, 2019) As seen in the mentioned data, pancreatic cancer has a very poor prognosis.

A previous study mostly related to our project, potentially useful drugs were identified using Connectivity Map (CMap)-based gene expression correlation assays. An algorithm (Met-express) was then applied to predict key pancreatic cancer enzymes involved in pancreatic cancer metabolism (Ma et al, 2016).

Another study explains the Met-express method, which includes the cancer gene co-expression network and the integration of the metabolic network (Chen et al, 2013).

Another related study, the LINCS L1000 project is a new, low-cost, high-throughput new version of the CMap project, in which drugs and disease states are linked through common gene expressions (Subramanian et al., 2017).

Our main motivation is to develop a new combined approach to predict candidate drugs for pancreatic cancer treatment based on the prediction of key enzyme-coding genes using machine learning algorithms and computational methods, thus we would change poor prognosis of pancreas cancer.

1.4 Goal / Contribution

Prediction drugs from correlations between differentially expressed (DE) genes of drug treated cell lines and cancer-causing DE genes is a conventional drug repositioning method. However, since the negative correlation observed between drug-induced and disease-specific changes in gene expression may not be always biologically significant, different methods are needed.

The aims of this project are to create gene co-expression networks after preparing the data with appropriate pre-processing methods, to divide these networks into modules, for predicting key enzyme-coding genes with machine learning algorithms and different

computational methods, and to computationally verify the obtained predictions. Thus, candidate drugs for pancreatic cancer treatment are predicted by predicting key enzyme-coding genes.

Unlike the traditional approach, the proposed method; The observed negative correlation between drug-induced and disease cell-specific changes in gene expression may not be biologically significant. Therefore, additional studies are needed.

1.5 Project Scope

Several methods will be done for prediction of a new therapeutic drug for treating pancreatic cancer by using machine learning models on gene networks, so the network should be generated. This network will partitioned properly to be used in a machine learning model.

The first step is preparing the data. This includes data pre-processing, generating the gene-network according to the pancreatic cancer cell data and the drug-cell relationship data and network splitting.

The second step is constructing the model. This includes the developing machine learning algorithm, validating and testing steps. The model will be asked to generate a prediction for a new possible drug unless any tribulation occurs.

The third method is testing the predicted value in a computational way to check the accurateness of the whole process.

1.6 Methodology / Tools / Libraries

The needed data sets to provide to the machine learning model after processing will be obtained from open-source databases. The data sets will be containing microarray gene expression data for both tumor and healthy (control) cells. R programming language will be used during the whole process of building a machine learning algorithm to predict a therapeutic drug as it provides many benefits with the *limma* library to process the microarrays and capability to the easiest data handling, modelling and also its flexibility and readability.

In order to check the negative correlation between our gene expression data set and the drug-induced gene expression profiles, the drug samples are needed. These samples will be obtained

from LINCS L1000 (Subramanian et al., 2017) which covers drug-induced gene expression profiles of several cancer cell lines.

Cancer gene co-expression network should be constructed according to the cancer gene expression dataset. This network is going to be compared with a metabolic network to build a model for prediction of enzyme-coding genes which has a role as a metabolite in pancreatic cancer cells. For this purpose, KEGG database will be used which has gene enzyme information. This step will bring the light on to the new enzymes which have a possibility to be used to cure pancreatic cancer.

The project will be presented on a web page via the Web API. ASP.NET was chosen as the framework. Microsoft Visual Studio will be used as the IDE and C# as the programming language.

CHAPTER TWO

LITERATURE REVIEW

2.1 Related Works

2.1.1 Prediction of Candidate Drugs for Treating Pancreatic Cancer by Using a Combined Approach

This study has several intersections with our thesis. The combined approach, which is used in this study, starts with collecting the datasets from NCBI GEO. Samples in a pancreatic gene expression dataset from tumor and normal, instead of non-tumor, tissues of the same person are used. These normal samples taken from several regions such as peripheral blood mononuclear cells and saliva. The limma package of R and Bioconductor has been used to identify the differentially expressed (DE) genes between normal and cancer tissues. Then, correlation between DE genes and the drug-induced ranked gene lists in Connectivity Map (CMap) (Lamb et al., 2006) is determined with respect to the Kolmogorov-Smirnov test (Hollender et al., 2013). This test brings the light on to the up-regulated genes in pancreatic cancer which is down-regulated in the drug-treated cultured cells or vice versa. AI and ML techniques are undeniably powerful in drug prediction as we know so far. Met-express procedure, which is an ML algorithm, is applied to the samples for prediction of key enzyme-coding genes; the results are validated by functional enrichment analyses and literature reviews. KPC enzymes are used to select candidate drugs. PUDs are considered to be used as a candidate drug for pancreatic cancer if its target is KPC enzyme or it is the substrate or product of the enzymatic reaction catalyzed by a KPC enzyme. The PUDs' targets are obtained from the DrugBank database to find the drug candidates. These predicted drugs are used for experimental validation (Ma et al., 2016). As Progesterone has been validated as an inhibitor for pancreatic cancer and also has a protective role for ovarian and endometrial cancers (Kim et al., 2013), this study can be evaluated as a successful prediction of a candidate drug for pancreatic cancer but with the newer version of CMap, which is LINSC L-1000 (Subramanian et al., 2017), would have been offering a better prediction.

2.1.2 High-throughput gene expression analysis for drug discovery

One of the studies has been made by Gregory GLennon to analyze the gene expression datasets with high throughput (GLennon, 2000). He mentioned some of the techniques has been used in this field which mainly focused on measuring mRNA expression levels for individual, well-characterized genes, or use in vitro nuclear 'run-on' transcription assays but it was also said that these techniques are inadequate as they use the transcriptional profiles of several active genes simultaneously. Hence, high-throughput screening (HTS) methods have been developed to be able to conduct large-scale screening and developing expression profiles patterns for tissues or cells. These methods provide identifying of the expression levels of novel genes and characterizing them, correlating mRNA expression patterns in many tissue types with disease states, identifying side effects of current and experimental treatments, and determining the effects of compounds on non-target tissues. Certainly, the most important method is DNA microarray. In this issue, DNA microarrays are defined as measurement of expression by using templates containing a lot of probes that are exposed simultaneously to a target sample, so they make it possible to survey DNA and RNA variations for drug discovery and evaluation. Then, GLennon continues using these obtained samples to prioritize drug targets and lead compounds. The approach is to compare the samples with a large gene expression database. In this direction, the GeneExpress database (BioExpress™) is used for target selection and prioritization. Searches are performed to find the genes that are downregulated with a sample or a set of samples. These searches provide the data for drug targets that counter differential regulation in diseased tissues. In addition to that, this database is mentioned to be used for determining whether a group of genes is associated with any disease which might be helpful for finding a drug target from the existing drugs. In conclusion, this study shows usability of microarray datasets by comparing it with other methods and how to use them properly for gene expression analysis for drug discovery successfully.

2.1.3 Identifying Drug Targets in Pancreatic Ductal Adenocarcinoma Through Machine Learning, Analyzing Biomolecular Networks, and Structural Modeling

This work includes the motivation to find new genes related to this disease and identify their targets for the diagnosis and treatment of pancreatic ductal adenocarcinoma (PDAC) (Yan et al., 2020). The study emphasizes the importance of having knowledge at both the systems and molecular level to develop PDAC therapy. In this study, briefly, a method developed based on predicting cancer genes and targets of potential drugs to prevent cancer was developed using protein-protein interaction (PPI) data and unrelated expression microarray datasets. The datasets used for this method included samples from cancer tissue and normal pancreatic tissue samples. The normalized data were analyzed to determine differentially expressed genes by the Benjamini-Hochberg method. As the information from PPI networks is not sufficient to identify disease genes and drug targets, a three-stage pipeline was developed. In the first stage, DEGs were sorted and the most relevant features were selected with the Recursive Feature Elimination (SVM-RFE) algorithm (Guyon et al., 2002). In the second step, a PPI network of DEGs was created with the STRING database. In the last step, a new score (RN) was defined for each gene. After constructing the PPI network, two analyzes were performed, including the calculation of the betweenness and closeness parameters . Then, the interactions between molecules and potential drugs were investigated with different analysis methods. The resulting metrics showed that RNs could be used to identify new disease genes and drug targets.

2.1.4 Prediction and identification of synergistic compound combinations against pancreatic cancer cells

Gemcitabine, alone or in combination with other drugs, is mostly used as the first line of treatment in patients with adenocarcinoma in whom the pancreas cannot be removed from the body, but the effect of gemcitabine is very low. With this motivation, this study identified synergistic combinations of compounds against pancreatic cancer cells with a computational method developed to select synergistic combinations of compounds based on transcriptomic profiles by both disease and compound. First, path signatures were used instead of the gene signatures of the compounds because they were more robust. A new hypothesis was then put forward that the pathway irregularities

potentially lead to compound synergy. It was then hypothesized that modifying compounds could be identified in PANC-1 cells when the disordered pathways of the disease were targeted. After these assumptions, two score types, Score1 and Score2, and the paths that are important in PANC-1 were defined later, and the Res-score was created. Based on these three scores, it was experimentally confirmed that 30 candidate compounds can be used in combination with gemcitabine (KalantarMotamedi et al., 2021).

2.2 Algorithms

2.2.1 Significance Analysis of Microarray (SAM)

SAM is used for large-scaled microarray data of gene or protein. It applies a t-test for each individual gene or protein level to determine if the expression is significant or not for the gene or protein, so it provides a better solution for analyzing large-scaled data than p-value cut-off of 0.01 which identifies 100 genes/protein out of 10,000 by chance (Iacobuzio-Donahue. et al., 2003).

SAM computes significance based on the standard deviation of the observation from the expected value; the expected value is computed by performing several permutations of the original samples (Petrisis, 2018).

2.2.2 Quantitative Structure Activity Relationship (QSAR)

QSAR has been applied to bioinformatics for drug discovery as it provides identification of chemical structures with good inhibitory effects on targets and low toxicity. QSAR studies aim to find a mathematical relationship between the activity and one or more descriptive parameters related to the structure of the given molecule (Abdel-Ilah et al., 2017).

In QSAR, the structure of a molecule should cover the features related with its physical, chemical and biological activities. First, the 3D models of molecular structures are needed to generate the molecular structure descriptors. After that, feature selection methods are applied to obtain the most important descriptors. The model is constructed with the selected descriptor sets and validated by predicting the activity of compounds in the external prediction set.

- IN (Veljovic et al., 2017), two multilayer feed forward neural networks and docking studies were developed to find out the hypothetical binding mode of the target compounds.
- Two 3D-QSAR models were designed for a series of non-purine xanthine oxidase inhibitors to study different factors affect the oxidase inhibitors (Li, 2013).
- A three-dimensional QSAR study has been implemented to study epothilones – tubulin depolymerization inhibitors (Lee et al., 2001).

2.2.3 Met-express

This algorithm was constructed for predicting the key enzyme-coding genes (Chen et al., 2013). First, Pearson Correlation Coefficients (PCC) are calculated for each expression value of all paired genes. The gene co-expression network was built according to the top three genes with the highest PCC for a given gene. Then, the partitioning method Qcut was used to divide the network into gene co-expression modules. The true positive (cancer samples) and false positive (normal samples) quantities have been counted for every median of gene expression value. ROC curve was plotted for the corresponding TPR against FPR for every threshold. If the median expression values of a cluster in cancer samples are greater than those in normal samples, then the ROC curve will be above the diagonal.

Metabolic network reconstruction was made according to the previous free version of the human KEGG Markup Language (KGML) files that contain enzymatic reactions, associated metabolites, and gene enzyme information from the KEGG database (http://www.genome.jp/kegg/). Enzymes are linked to their shared metabolites. For a given enzyme-coding gene, Met-express computes an importance score by the following equation:

$$Score_i = |AUC_{ROC} - 0.5|(\frac{C_{in}/C_{all}}{N_{in}/N_{all}})$$

where the first term |AUCROC − 0.5| represents the cancer specificity of a gene co-expression module, and the second term indicates the enrichment degree of metabolic links of this gene within the module. C_{in} and C_{all} are the quantities of enzyme-coding

genes connected from the test gene inside the module and inside the whole network according to the metabolic network respectively. N_{in} and N_{all} are the quantities of enzyme-coding genes in the module and in the metabolic network for each background respectively.

2.2.4 SVM-RFE Algorithm

Support Vector Machine (SVM) is a machine learning algorithm. It is mainly used in solving classification problems. (Ray, 2017). It is used for feature selection as well as classification.

SVMs can analyze gene expression patterns from DNA microarray data. Even if the number of patterns is small, they can handle many features.

SVM-RFE works by measuring the weights of features with respect to their support vectors, but noise and non-informative variables in high-dimensional data can affect the hibernation of the learning model. Therefore, there is a mutual information (MI)-SVM-RFE method that executes SVM-RFE to select distinctive features after filtering out noise and non-informative variables.

The RFE method allows finding subsets of genes to find the optimum number of genes. It is also convenient to use this method for linear classifiers (Lin et al., 2012).

2.3 Tools and Libraries

2.3.1 limma Package

Limma is a library in the R programming language. It is used for expression analysis of microarray data as well as for analysis of RNA-seq data. limma can work with all gene expression technologies.

2.3.2 Scikit Learn

Scikit-Learn is a library in the Python programming language. Developed for Machine Learning studies and applications. It can be used in both supervised machine learning and unsupervised machine learning.

2.3.3 Cluster Profiler package

This package is used to analyze and visualize functional profiles of genomic regions and genes. clusterProfile automates the classification process of biological terms and enrichment analysis of gene clusters. Currently clusterProfiler supports three species: human, mice, and yeast (Yu et al., 2012).

2.3.4 KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database used for the analysis of gene functions in terms of metabolic and signaling pathways. It contains a large number of different data such as diseases and drugs that can be used in the research areas of health and bioinformatics science.

2.3.5 Web API

Web APIs can be defined as programmable interfaces. They are used in mobile and desktop applications or browsers on different devices such as mobile phones, computers. They enable data exchange between clients and devices.

2.3.6 Microsoft Visual Studio

Visual Studio is an IDE developed by Microsoft. eNot developed for just one programming language or tool. It also supports different programming languages. It can also be used when developing websites, web applications or for mobile development purposes.

CHAPTER THREE

REQUIREMENTS / REQUIREMENT ENGINEERING

3.1 Functional Requirements

3.1.1 Determining the Differentially Expressed Genes

User's and the available microarray data will be examined to determine the differentially expressed genes. The genes will be pre-selected according to their expression levels and network ranks to provide an input score for the protein-protein interaction (PPI) network.

3.1.2 PPI Network Analysis

The network will be used to prioritize the genes with the help of subnet enrichment techniques. The prioritized genes will be used as biomarkers for further analysis.

3.1.3 Drug Searching

Drug databases will be scanned to find binding compounds for the prioritized genes. The results will be validated whether the candidate drug has been used in clinical trials for treatment of pancreas cancer. The validated drugs will be the output of the system.

3.1.4 User Interface

Users will be able to upload their microarray dataset to be analyzed by the system and they will obtain a candidate drug to be used for treatment.

3.2 Non-Functional Requirements

3.2.1 Reliability

Results at the end of the project should be acceptable, logical, and worthy of review in the wet-lab or already under phase trials. Compounds that emerge at the end of our project will be verified by checking from PubChem and Clinical Trials databases.

3.2.2 Availability

The web APIs, web server and domain to be used should always be largely accessible from the user at all times.

3.2.3 Performance

Although our work does not have commercial concerns about customer satisfaction, the loading speed of the platform to be prepared to save time for users in their work and the speed of arrival of the results are very important. 4GB RAM, 1024 x 768 resolution Super VGA is required for effective performance. Browsers should be Microsoft Edge running on Windows 11, Windows 10 and Windows 8.1, Google Chrome running on Windows 11, Windows 10, Windows 8.1, Apple Safari running on the last two versions of Mac OS.

3.2.4 Usability

As a result of our work, we aim to develop a website where users can find predictions of candidate drugs for pancreatic cancer by uploading their own gene expression data. This website should be very comfortable to use for end users. In addition, this website should contain the necessary explanations for the purpose of use.

3.2.5 Security

The website to be built should be resistant to possible attacks. In addition, data to be uploaded by users must not be lost or stolen.

3.2.6 Maintainability

First of all, the algorithms used over time should be made more effective. In addition, owing to increasing data, the number of drugs that can be accessed will increase. In addition, various components of the site, such as the interface, should be regularly maintained. In addition, the algorithms should be constantly tested and the detected errors should be corrected.

3.2.7 Scalability

When data volume and the number of users of the system increase, it should not be harmed by this situation and should continue to work.

CHAPTER FOUR

DESIGN

4.1. Architectural View

As seen in the Figure 4.1, our system starts with the obtained gene expression data of normal and cancer samples of patients from the NCBI GEO database. This data consists Affymetrix Human Genome U133 Plus 2.0 arrays. According to these data obtained; data normalization, filtering ambiguous data, statistical significance (e.g., t-test, two tails, unequal variation) and fold-change calculations, and assigning probe set as the gene's expression level operations are applied.

Then, using these data, the data provided by the Bioconductor and STRING databases, a protein-protein interaction (PPI) network is created.

After a PPI network is created, pre-selection of hub genes, construction of PPI network according to the selected hub genes, gene prioritization with subnet enrichment techniques, drug exploration methods are applied. Drugs were discovered and validated using the Drug Gene Interaction Database. Finally, apart from the results we found in our study, there is a service website in which users can also upload new gene expression data to obtain new treatment options.

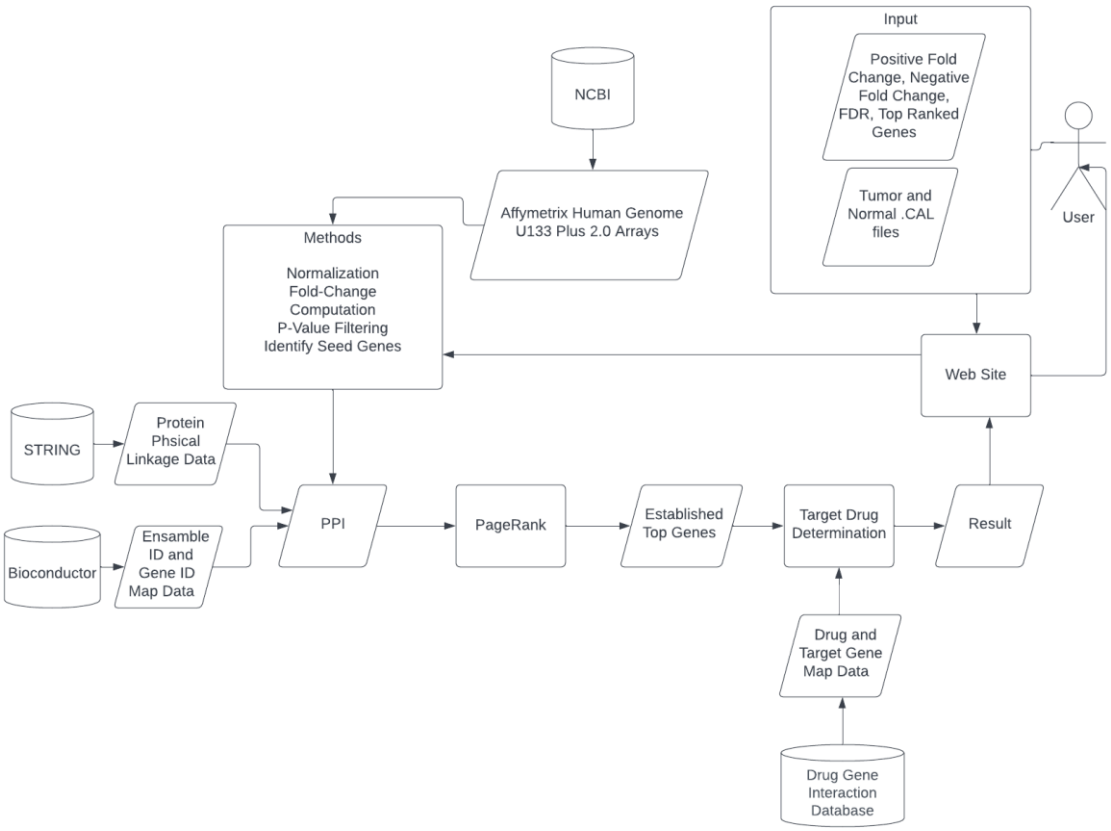

Figure 4.1 Architectural View

4.2 UI Design

The website contains sections where users can access brief information about us and our article. The main function of the site is that users can access treatment suggestions by uploading their own data. The upload function is on the main page (Figure 4.2.1). After users upload the expression data to the system, the results are generated as seen in Figure 4.2.2. The results provide information to users in five different columns: drug name, brief information about the drug, compound structure, clinical trials and drug targets.

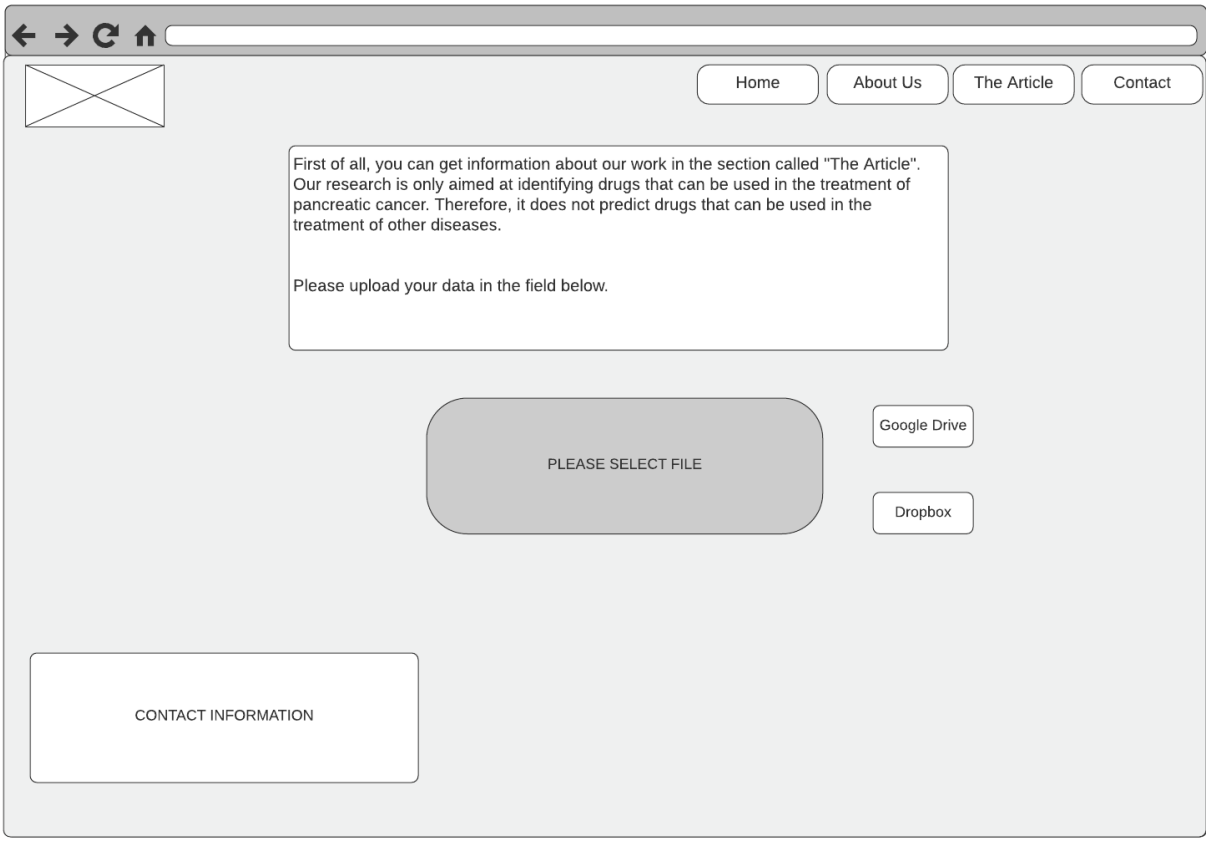

Figure 4.2.1 UI Design

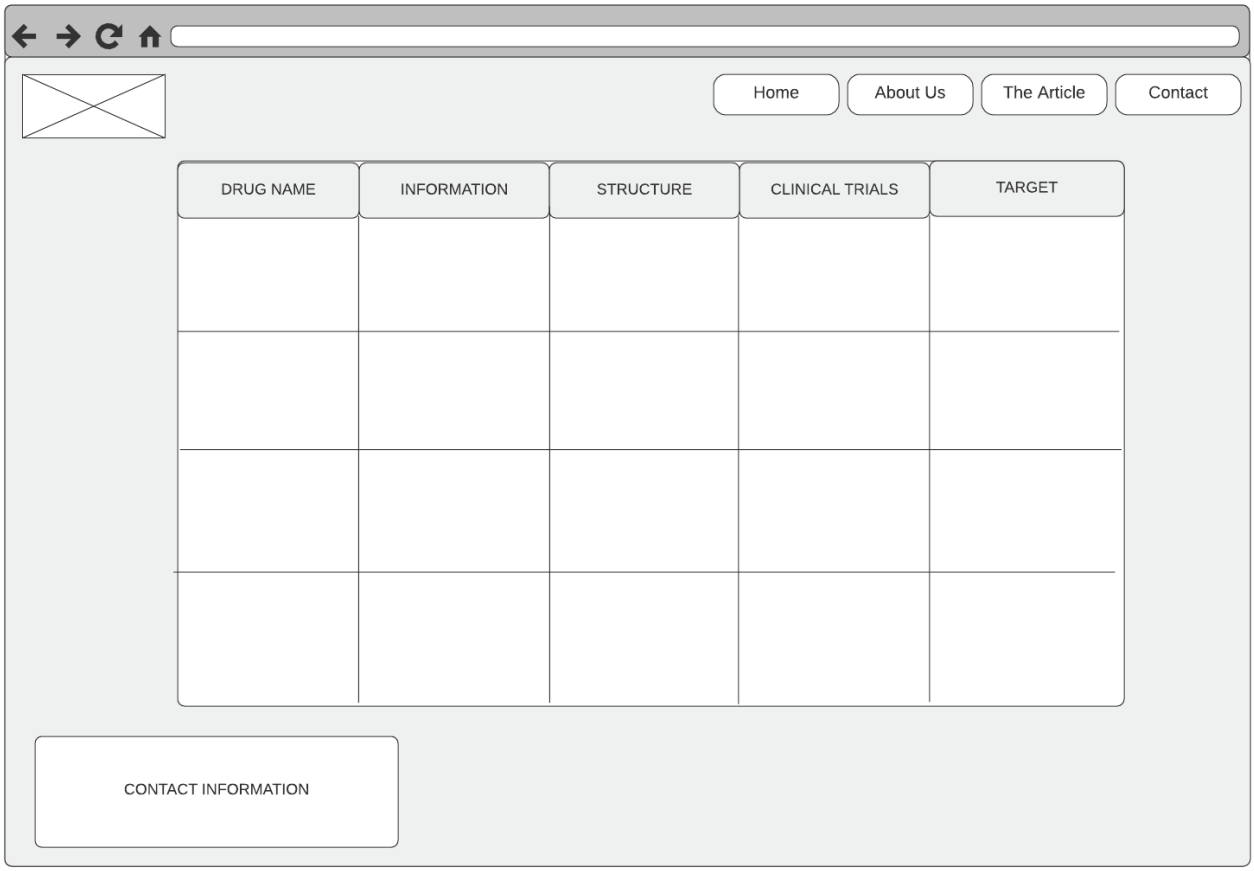

Figure 4.2.2 UI Design

4.3 Class Diagram

As seen in the Figure 4.3, the *DataCollector* class is interacting with other classes to reach all the databases through their corresponding API's. After the input is fed in the *User* class and all data is obtained from the databases, the *ExpressionAnalysis* class is being aggregated to handle the preprocessing and differential expression analysis. The differentially expressed genes are transferred to the *NetworkAnalysis* class to perform pre-selection and subnet enrichment on the PPI network. The prioritized genes are passed to the *DrugExploration* class to obtain drug candidates. These drug candidates are used in the *Validation* class to check the accuracy of the predicted therapeutic drug.

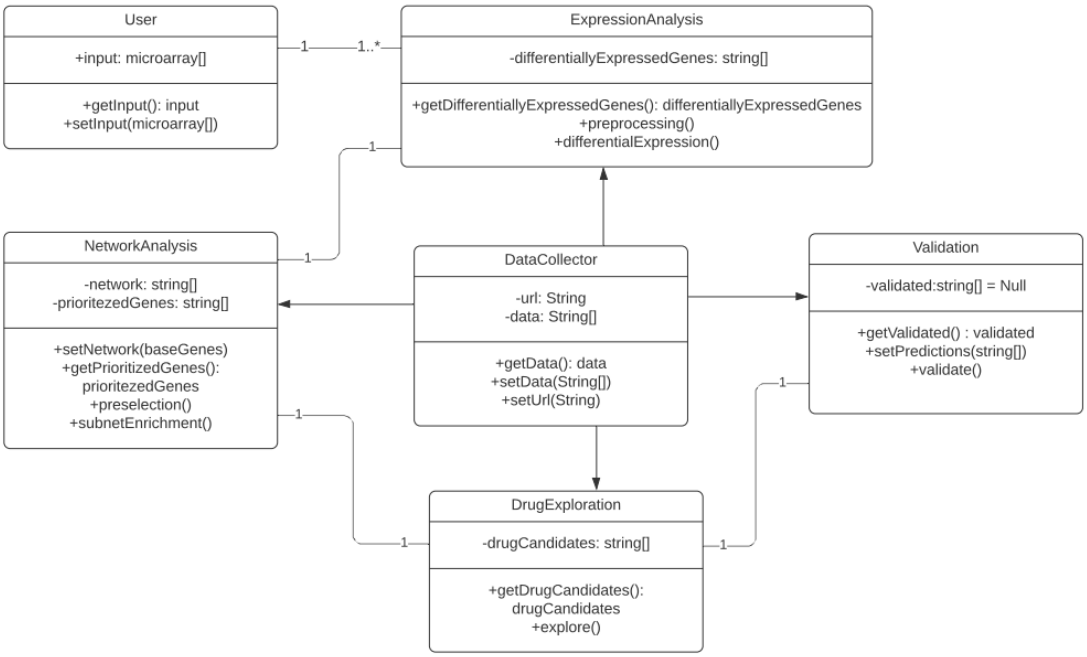

Figure 4.3 Class Diagram

4.4 Use Cases

The main actor of the system is the user as shown in Figure 4.4. User provides an input to get a prediction result of the system. This input is passed to the pipeline, in case it is valid. Otherwise, it may be rejected if it is not given in the correct form. At the end, users can monitor the output which has been produced by the system.

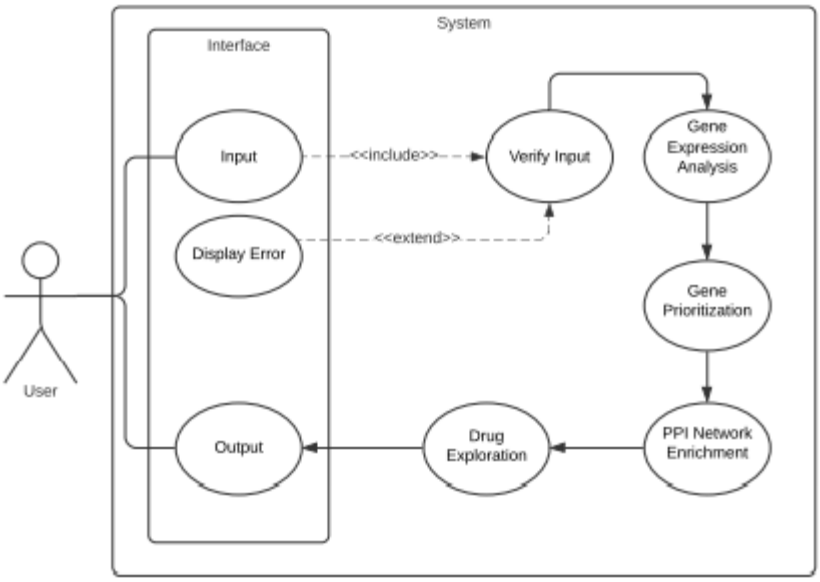

Figure 4.4 Use Case Diagram

4.5 Sequence Diagram

First, users need to provide the input to be validated by the system (Figure 4.5). If it is valid, then it is being passed to the pipeline with other data from the databases relatedly. Pipeline does include all the required functionalities and starts to be executed when the input and required data are obtained. All of the external data will be collected from databases through their provided APIs. Data will be manipulated to get insights by the following steps in the Pipeline; gene expression analysis, gene prioritization, network subnet enrichment and drug exploration alternately. In the end, the output will be represented to the user.

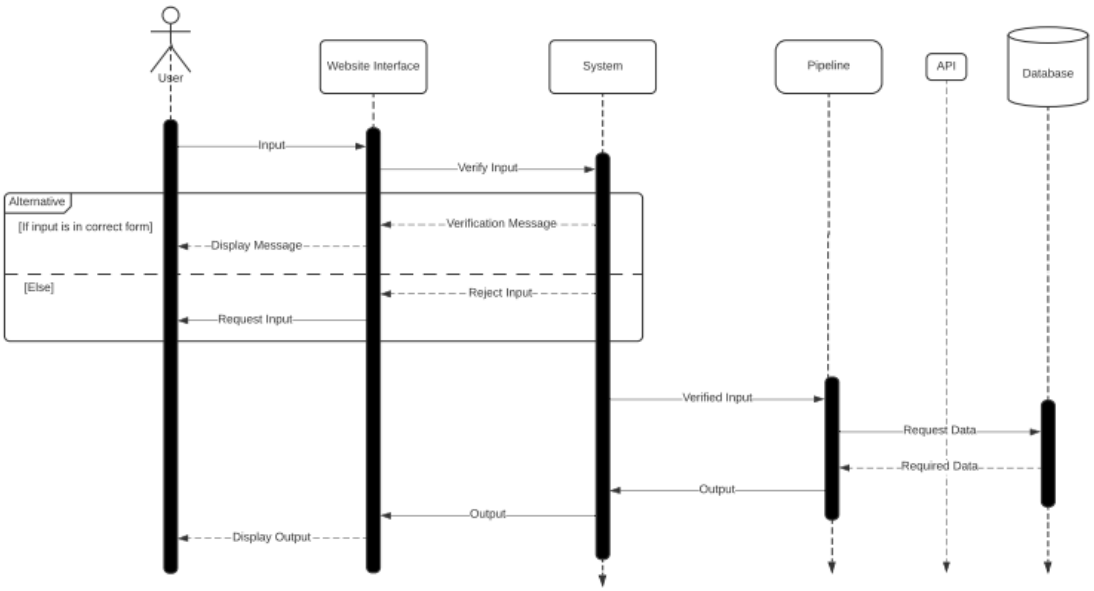

Figure 4.5 Sequence Diagram

4.6 Activity Diagram

Users reach the system via a browser if it is supported as shown in Figure 4.6. Then, an input will be asked in a specific form. As a result, users will be informed of their possible treatment candidates corresponding to their input.

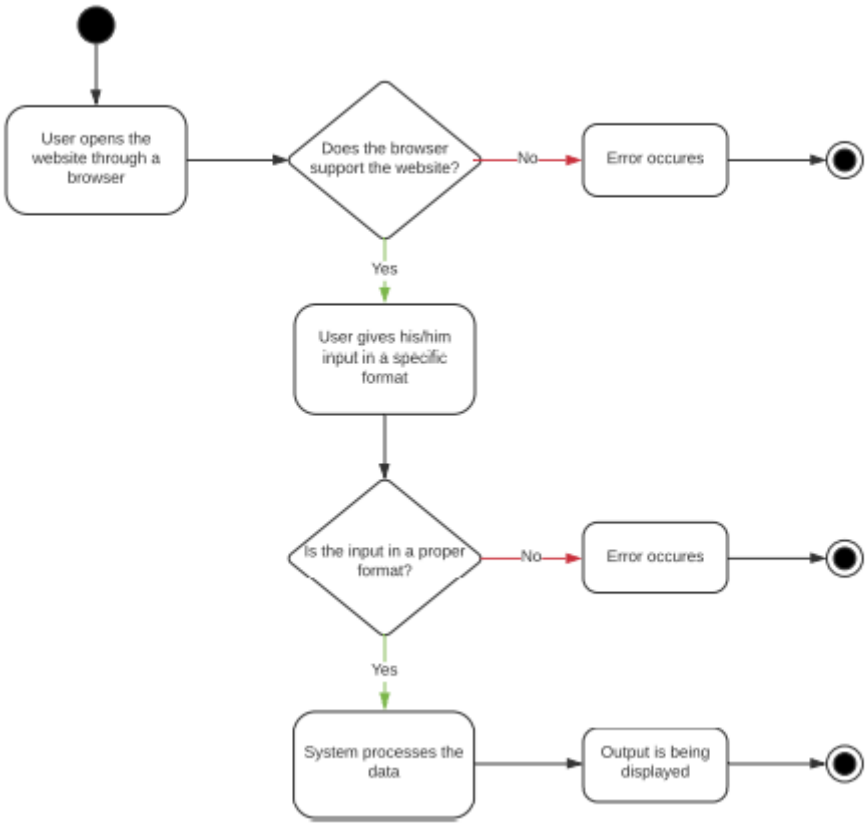

Figure 4.6 Activity Diagram

4.7 Deployment Diagram

The deployment diagram in Figure 4.7 shows the components in which the platforms used in the project are summarized. A Load Browser is used, because more than one API server is used. It is intended to distribute the workload using the Load Browser. Finally, a connection is made with the browser using HTTPS, user requests are accepted and verified responses are sent to the user.

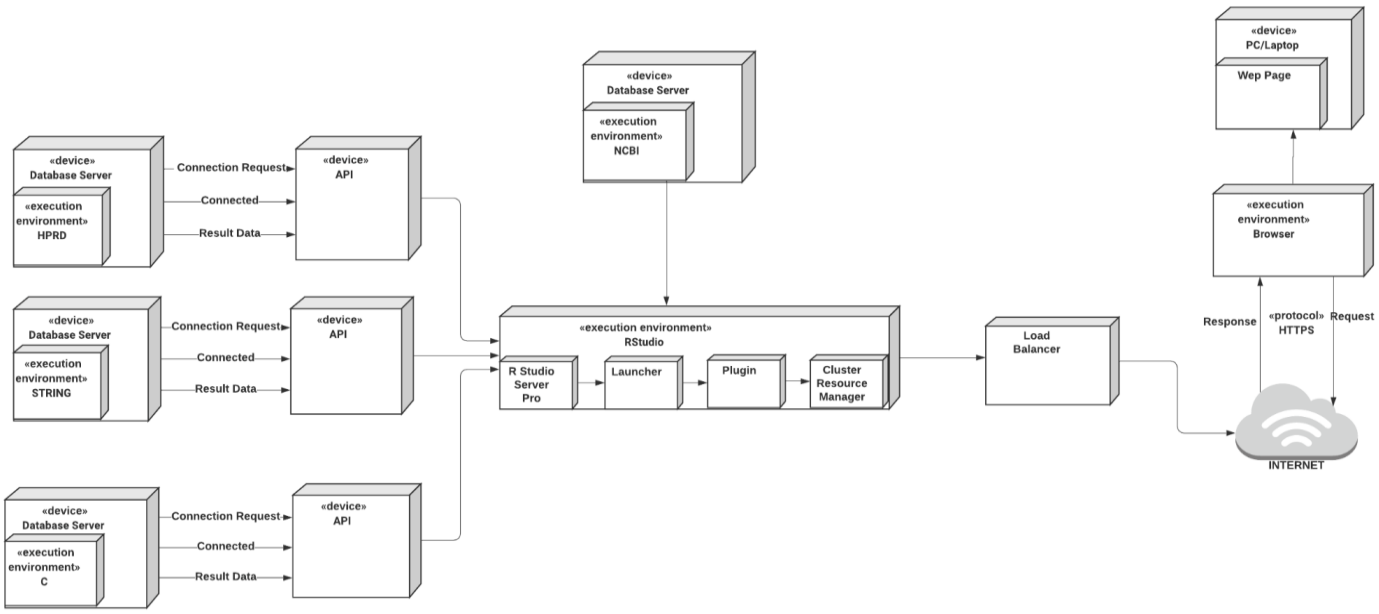

Figure 4.7 Deployment Diagram

CHAPTER FIVE

IMPLEMENTATION

5.1. Microarray Data Preparation

The microarray data to be used in this study have been taken from the public NCBI GEO datasets. 72 (36 normal and 36 tumor) tissue samples, 32 (16 normal and 16 tumor) tissue samples have been taken from the datasets coded as GSE15471 and GSE16515 respectively and each dataset was analyzed severally. Expression values were normalized and background corrected with the RMA method.

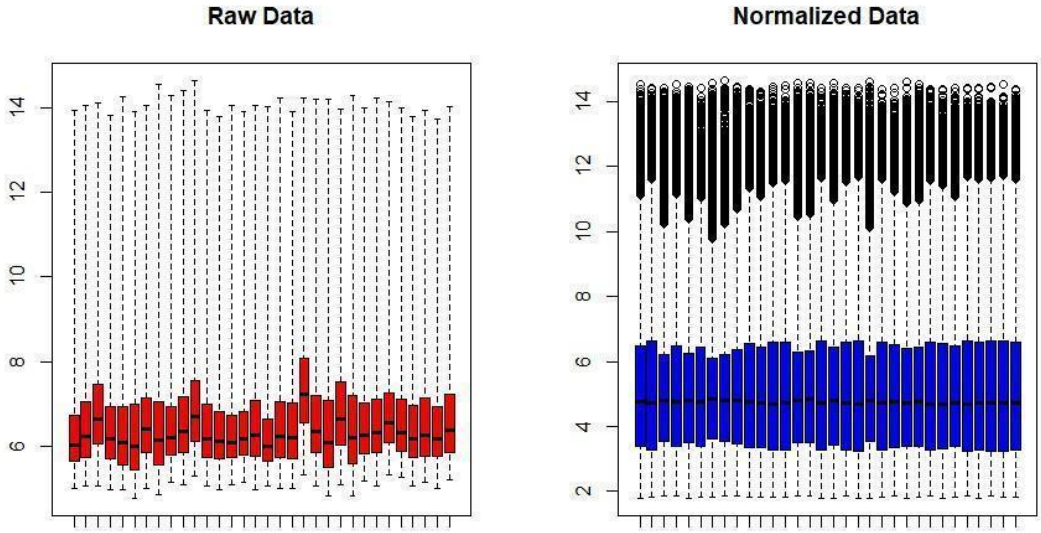

Figure 5.1 Box Plot of Raw (left) and Normalized Data (right) for GSE16515

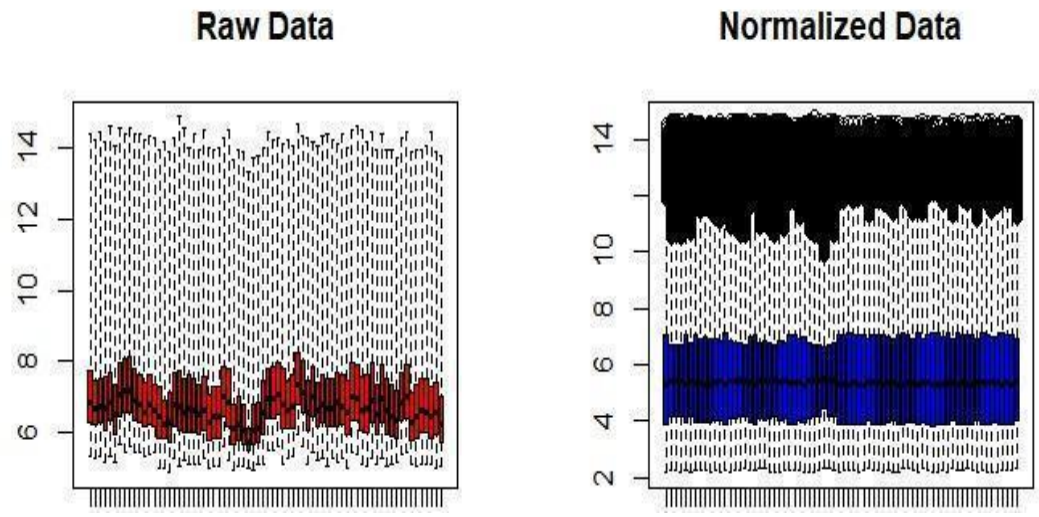

Figure 5.2 Box Plot of Raw (left) and Normalized Data (right) for GSE15471

As seen from Figure 5.1 and Figure 5.2, the medians of each sample have been equalized by applying normalization. This way, the difference between the tumor and normal samples, which is known as fold change, will not be dominated and deflected by any sample.

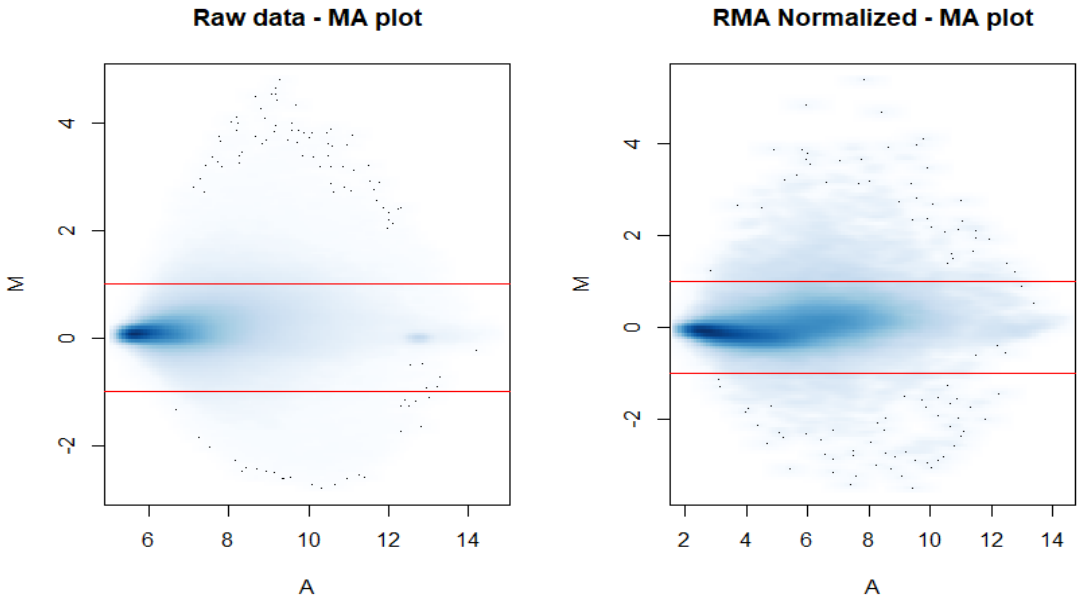

Figure 5.3 MA Plot of Raw and Normalized Data for GSE16515

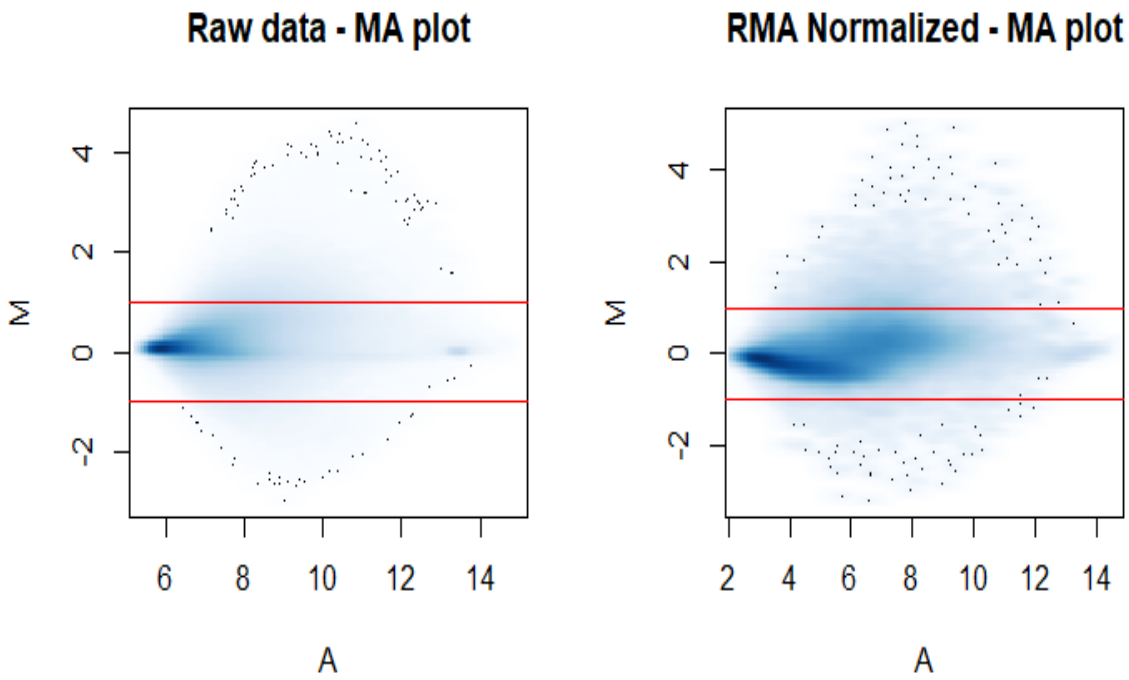

Figure 5.4 MA Plot of Raw and Normalized Data for GSE15471

The normalization made more genes to be appeared above and below the threshold value as seen in Figure 5.3 and Figure 5.4.

The normalized data were mapped with gene name annotations according to their probe identifiers. After that, we realized some probe identifiers are linked to the same genes. For this reason, the data were aggregated under their gene annotations and performed a mean calculation per gene. Thus, each tuple with the length of 20862 was annotated with a unique gene. For each gene, a fold change (FC) value was calculated by subtracting the expression levels of normal samples from tumor samples. In addition to the fold change values, *t*-test has been applied to each tuple and observed their corresponding *p*-values. As the fold change value remarks the vectoral direction of differential expression levels, for a specific gene it is desired to be in the same direction in each dataset, so the genes which do not have the similar FC direction (i.e., all in up-regulated or all in down-regulated) among the two datasets were determined. After this control, 3835 inconsistent genes were removed, the following experiments were performed with the remaining genes.

The first experimental results are given in Table 5.1. Based on different threshold settings, the total number of differentially expressed genes (DEG) vary dramatically. The same threshold values should be set to be consistent in the rest of the analysis.

Experiment Thresholds		**Number of Observed Genes**	
P-Value	**Absolute-Fold Change**	**GSE16515**	**GSE15471**
0.01 <	1.0 >	402	878
	1.5 >	195	341
	2.0 >	85	127
0.05 <	1.0 >	605	902
	1.5 >	257	348
	2.0 >	103	130

Table 5.1 Experiment Results

5.2 APIs review and API testing

HPRD (Human Protein Reference Database): HPRD is a data set that contains very comprehensive and versatile information about human proteins. HPRD contains a wide variety of information about the interactions of human proteins with each other, protein inputs, and protein expressions. Contributing to HPRD data is possible. Data on human proteins can be enriched by loading them into this dataset. One of the datasets to be used in this project when constructing a PPI network is the HPRD dataset. A module called indra.sources.hprd is used to retrieve content from the HPRD dataset. This module is especially used to access the interactions of proteins with each other and to access protein complex interactions. Since INDRA is a Python package, this module will work with the Python programming language. Many of the features of the INDRA module are available via the REST API. The local address

"http://api.indra.bio:8000" on the official website of INDRA has been tested on https://reqbin.com/. The API did not request token information during this test. As a result, 200 codes were returned from the tested API. This shows that the result was successful.

STRING Database: Another database that we will use while creating a PPI network is the STRING database, which also contains information on protein-protein interactions. In addition, the STRING database includes experimental data and computationally estimated data. The STRING database has an API that can be accessed with HTTP requests. We can access gene names and much more information with access numbers in this API. To reach the desired gene is possible with a URL format containing an identifier and optional parameter. This API has been tested on https://reqbin.com/ as stated. No token information was requested during this test. 200 code is returned from the tested API. This shows that the result was successful.

DrugBank: DrugBank is one of the databases we will use for drug discovery in our study. DrugBank is a fairly large drug database. The database contains data on drugs and drug interactions. DrugBank API is a REST API and uses HTTP methods. DrugBank API queries are filtered by region. Therefore, the region code in the API URL should be specified. To access the DrugBank API, obtaining an API key is necessary. DrugBank asks them to be contacted to get this API key on its official website. We contacted DrugBank to get the API key. We are waiting for a response in the current process.

STITCH Database: The STITCH database is another database to be used for drug discovery in our study. The STITCH database contains data on drug-target interactions. API access token is required to receive data from the Stitch API. The website at https://www.stitchdata.com/ provides this access token free of charge to its users who are members of it. In order to access the Stitch API, we became a member of this site and we obtained free API access token from the site. The API tested with this token on https://reqbin.com/. A successful result was obtained from the tested API.

Clinical Trials: Clinical Trials is the database we will use to validate the results of our research. This database includes the compounds studied, clinical phase I / II / III trials and tests performed. Clinical Trials provides an API service for us to access the data it contains. API data can be accessed via access query URLs or information URLs. The sample query URL generated by Clinical Trials has been tested via https://reqbin.com/. During this test, the API did not request token information. The result of the Clinical API test was also successful. In this way, it turns out that the Clinical API is also accessible.

5.3 User Interface

As a result of our work, we have created a web environment where users can upload their data so that users with different data sets can also benefit from our algorithm and experiment their own data. We used Node.js, Express.js, HTML, Css and JavaScript technologies while creating an interface where users can test their data. Visual Studio Code was used as the IDE.

The website we developed consists of a home page where users can upload their own data, a result page to see analysis results, an article page where they can review our work details, and a contact page where they can contact us.

5.3.1 Home Page

Home Page is where users can upload their files. Here we only allow them to upload microarray input files with .CEL extension, which is default data format for Affymetrix chips.

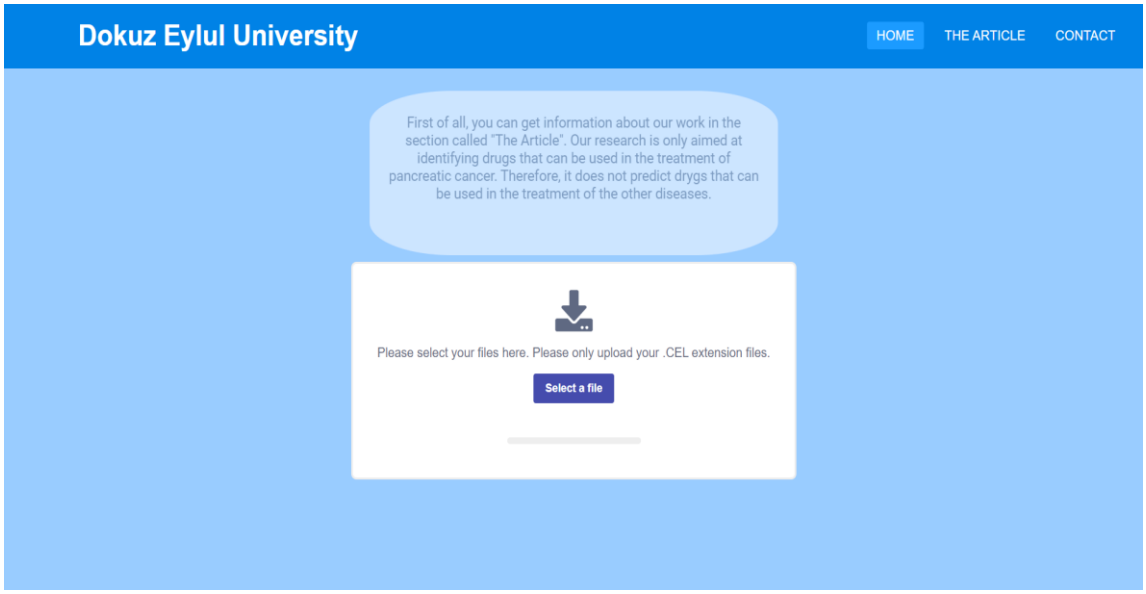

Figure 5.6 HomePage View

5.3.2 Article Page

The Article Page is the page where users can learn about the details of the work by reviewing all of our work. This page contains our project technical report.

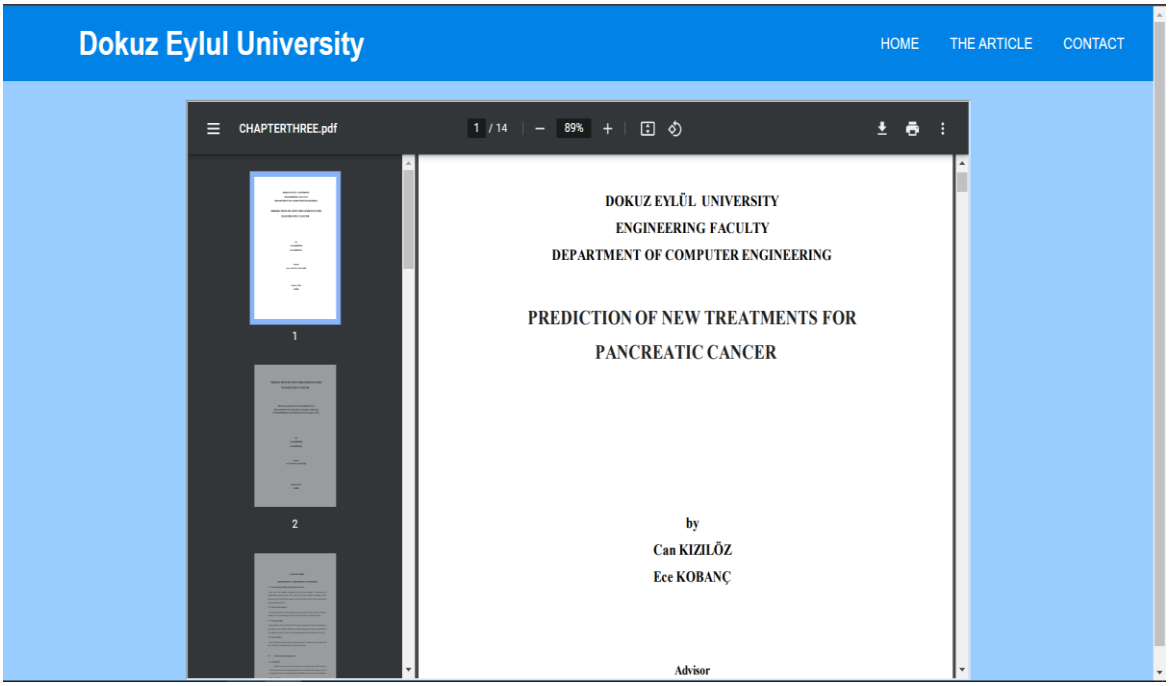

Figure 5.7 Article Page View

5.3.3 Contact Page

Contact Page is the area where users can contact us. Users can share their requests, opinions and criticisms with us using this page.

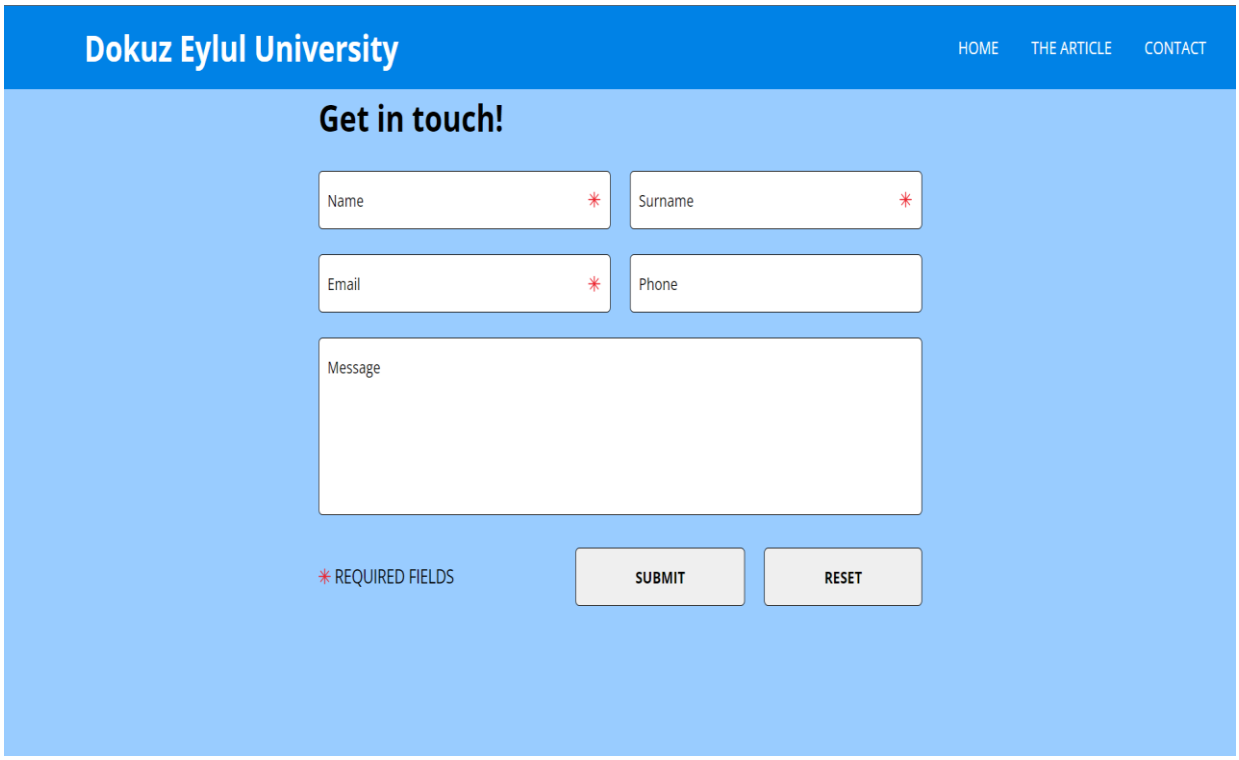

Figure 5.8 Contact Page View

5.3.4 Result Page

Result page displays the predicted drugs to the user as a result of the calculations made after uploading their data to the system. The results page also includes information about these compounds, the structure of the compounds, whether clinical trials of the predicted compounds have been conducted, and the targets of the drugs.

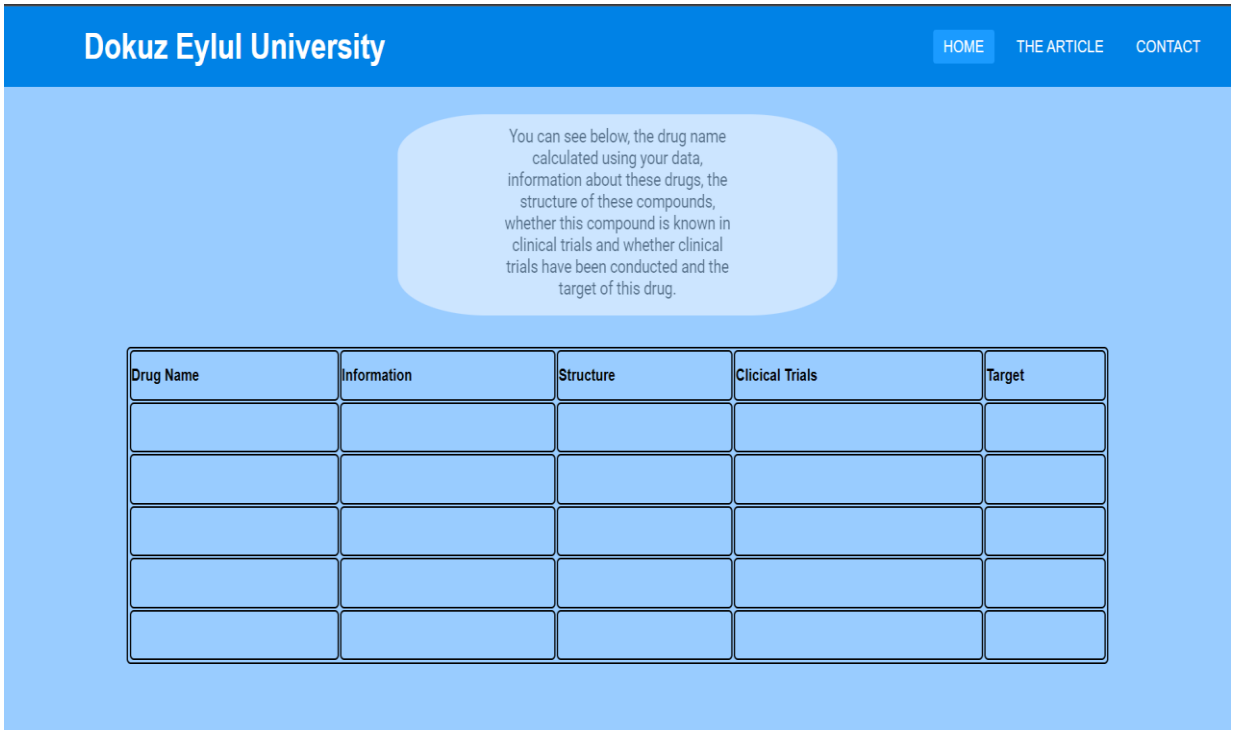

Figure 5.9 Result Page View

5.4 Network Data Preparation

Functional and physical protein interactions were obtained from the STRING database to construct a protein-protein interaction (PPI) network (*STRING Functional, 2022,* https://string-db.org/cgi/download?sessionId=bhrZZYPBLNst.). This dataset is an edge list of the PPI network to be created. The original STRING version 11 consists of 1.991.833 edges (interactions) for human proteins. The dataset contains three different attributes: "protein1", "protein2" and "score". First, proteins that have an interaction score (i.e., functional or physical interaction) less than 700 were removed from the original dataset. After this filtering, 164.123 interactions remained in the PPI network. The "protein1" and "protein2" are represented by Ensembl protein identifiers, which were converted to Entrez gene identifiers by using *EnsDb.Hsapiens.v86 library* in R-Bioconductor. After identifier transformation, these two columns are combined to form an edge list. The confidence score for each interaction is used as an edge weight when creating the PPI network. Proteins that have "NA" values as the Entrez gene identifiers were removed from the edge list. As a result of this operation, there are 10.044 vertices and 160.744 edges in the PPI network. The *igraph* library in the R was used to create the initial PPI network for human.

5.5 Page Rank

Determination of the most important genes within a PPI network was completed by running the with "*page rank"* function in the *igraph* library. Beside the network to work on, the page rank function uses one more parameter called *personalization vector*. This parameter defines the probabilistic distribution over the nodes in the network. The probabilistic distributions were obtained from the *p*-values of genes in each of the microarray dataset. As the lower *p*-value indicates the more chance to be selected while performing a random walk in the network, therefore *p*-values were exponentialized with minus one as given in the following equation:

$$\square = \left[\frac{1}{\square-\square\square\square\square\square_1}, \frac{1}{\square-\square\square\square\square\square_2}, \ldots, \frac{1}{\square-\square\square\square\square\square_\square}\right]$$

(*1)*

After that, each of the exponential *p*-values was divided by the summation of them (Equation 2). This way, the distribution was rescaled, in a way that it sums up to one.

$$Pn = \left[\frac{P}{\sum_{i=1}^{n} P_i}\right] \quad (2)$$

Finally, the *page rank* function was executed for each dataset separately with their corresponding *personalization vectors*.

5.6 Target Selection

The rDGIdb library was used for drug selection for genes selected from Dataset 16515 and Dataset 15471. rDGIdb queries DGIdb using R/Bioconductor.

This library accepts gene symbol as gene parameter when querying drugs that are targeting the given gene. All available source databases in the library were used to enlarge drug screening by the inquiry.

The available drug target databases are: *COSMIC, ChemblInteractions, CIViC, CancerCommons, ClearityFoundationBiomarkers, ClearityFoundationClinicalTrial, DTC, DoCM, DrugBank, FDA, JAX-CKB, MyCancerGenomeClinicalTrial, OncoKB, TALC, TTDharm, TALC, GuideToPharical, TTD, Medicine, GuideToPharical, TTD, Pharmacology, TCIGnicalTrial, TALC, NTC.*

Interaction type between gene and drug molecule is critical since it represents biochemical effect of a compound on the activity of target protein. Therefore we used filters made for selecting the accurate drug interactions. In the activating group of interactions, the given drug increases the biological activity or expression of the protein target. Therefore, drugs from this group were selected for genes have negative fold change value (down-regulated mRNA expression). Interaction types in the activating group are activator, agonist, chaperone, cofactor, inducer, partial agonist, positive modulator, stimulator, vaccine.

A drug in the inhibitory group reduces the biological activity or expression of the protein target. Therefore, drugs from this group were selected for genes with positive fold change values (up-regulated mRNA expression). Interaction types in the inhibitory group: *antagonist, antibody, antisense oligonucleotide, blocker, cleavage, inhibitor, inhibitory allosteric modulator, inverse agonist, negative modulator, partial antagonist, suppressor*.

CHAPTER SIX

EXPERIMENTAL RESULTS

The page rank method returns a ranking value for each node within a network. The higher rank values indicate more important proteins in the network. The top-hundred genes with highest rank values were chosen from the output of the page rank method. Selecting top hundred nodes value granted the enough number of genes, 16 genes are common between two microarray datasets.

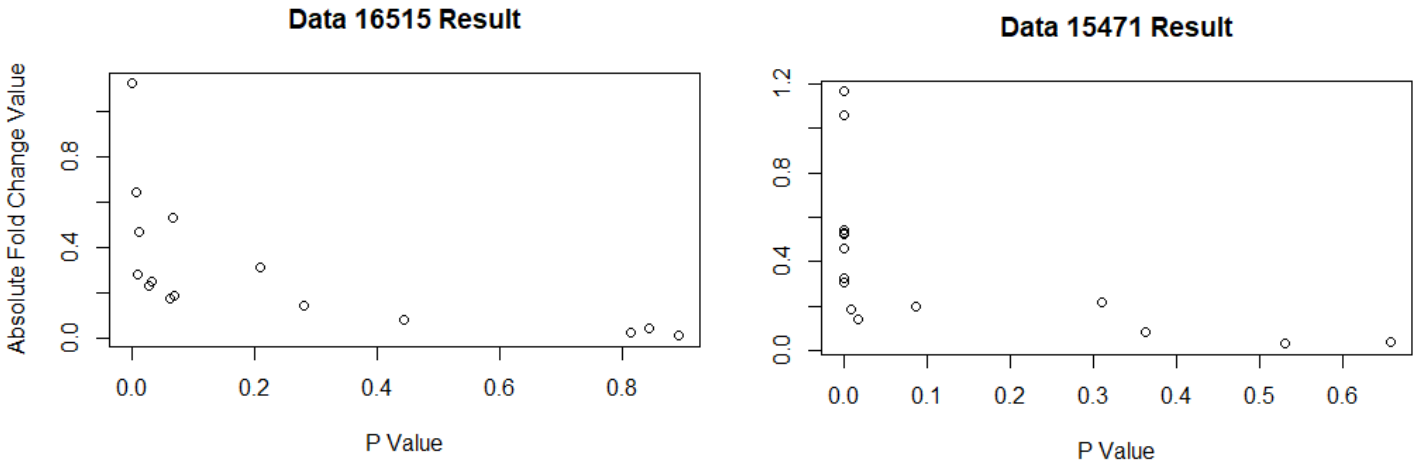

Figure 6.1 Significance analysis of the genes common for both datasets.

6.1 Dataset 16515

Based on the top-100 highest ranked genes, four of them were not involved in the actual microarray dataset. These are the genes with entrez identifiers as *135, 693196, 406974* and *100500810*. The rest of the genes were analyzed for their significance (*p*-value and fold-change) in the actual dataset as given in Figure 6.2.

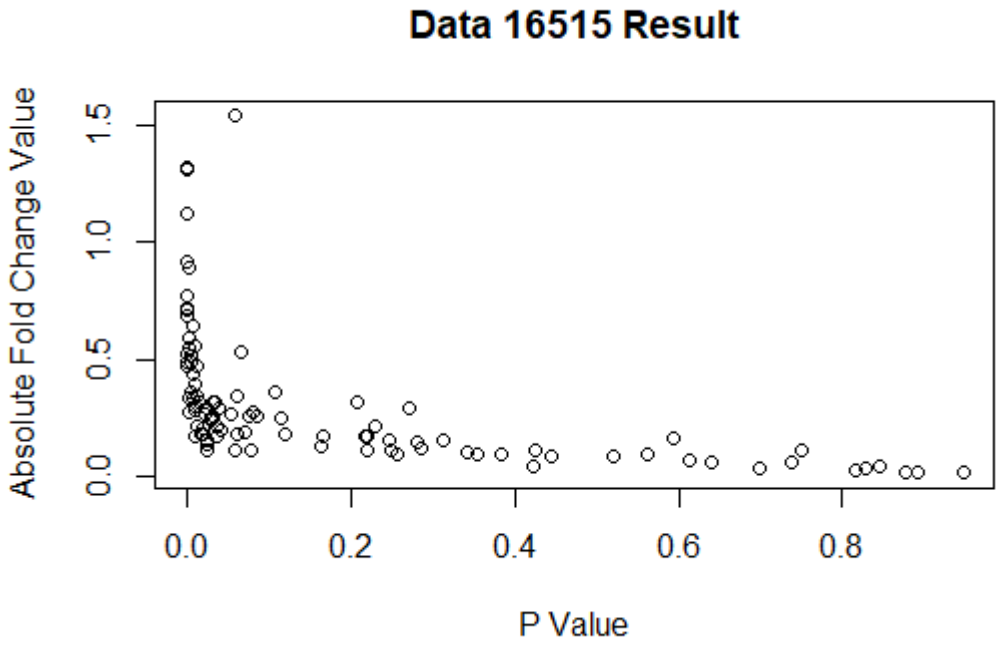

Figure 6.2 Significance analysis of the top-ranked genes in Dataset 16515.

6.2 Dataset 15471

Based on the top-100 highest ranked genes, five of them were not involved in the actual microarray dataset. These are the genes with entrez identifiers of *406974, 101928601, 100500847, 102465445* and *2790*. The rest of the genes were analyzed for their significance in the actual dataset as given in Figure 6.3.

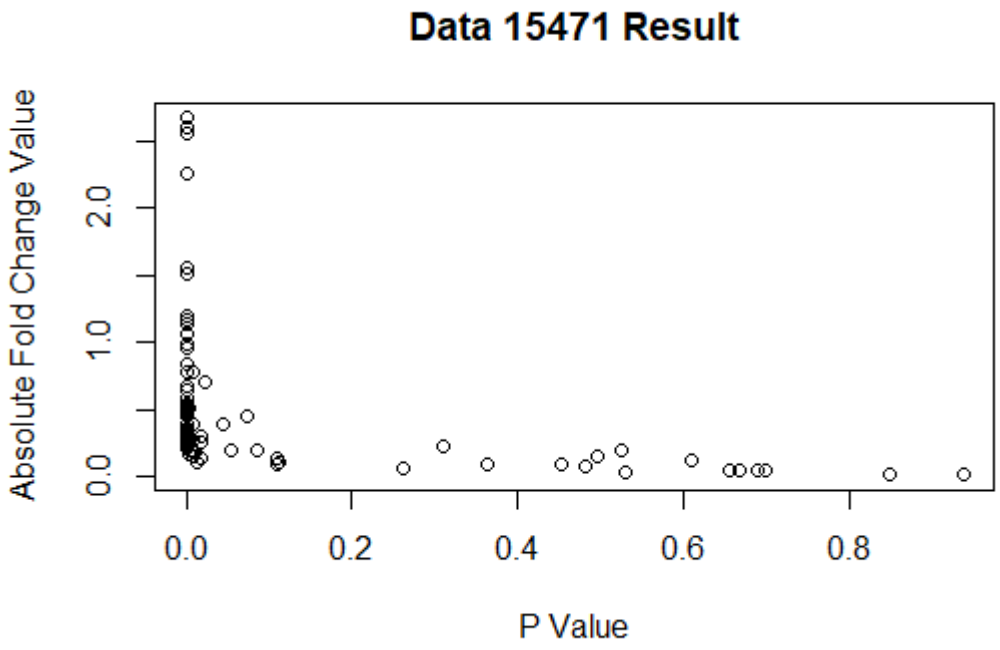

Figure 6.3 Significance Analysis of the Results of Data 15471

6.3 Target Selection

The genes in both datasets were divided into two as genes with negative fold change values and genes with positive fold change values. Thus, four groups were obtained that are listed in **Table 6.1, 6.2, 6.3, 6.4.**

Table 6.1 Genes with negative fold change values in dataset 16515.

Gene Identifier	**Fold Change**	**p Value**	**FDR**
PEBP1	-0.5878149	3.587364e-03	0.031537961
XBP1	-0.7086329	2.042241e-05	0.001589508
TUSC3	-0.5112467	4.460068e-03	0.035566907

Table 6.2 Genes with positive fold change in dataset 16515.

Gene Identifier	Fold Change	p Value	FDR
GNA15	1.1225389	3.404671e-04	0.0083073984
GNAI2	0.6450830	7.232028e-03	0.0467393360
HSP90AA1	0.7207327	2.426403e-05	0.0017761272
IL1RAP	1.3119918	1.986588e-05	0.0015698561
OXTR	0.8895206	2.968077e-03	0.0282353064
PKM	1.3215621	1.106675e-07	0.0001004229
ACBD3	0.7696696	8.050001e-04	0.0132548641
UBC	0.5223238	4.842018e-04	0.0099521360
YWHAG	0.5467531	3.073315e-03	0.0288808505
MARCO	0.9138011	1.663240e-03	0.0201970332
RAB11A	0.6821306	9.315163e-05	0.0037443726

Table 6.3 Genes with negative fold change in dataset 15471.

Gene Identifier	Fold Change	p Value	FDR
EGF	-2.2617520	5.187491e-05	1.369718e-04
F2	-0.6773190	3.716470e-07	2.669869e-06
PLCB1	-0.6752765	2.958577e-06	1.272878e-05
GNG3	-0.5053496	1.479889e-08	3.048931e-07
GNG7	-0.5134201	7.830076e-10	4.367675e-08
UTS2R	-0.5772072	1.578128e-08	3.174823e-07
S1PR5	-0.5785610	7.731424e-08	8.955746e-07

Table 6.4 Genes with positive fold change in dataset 15471.

Gene Identifier	**Fold Change**	**p Value**	**FDR**
RGS19	0.8447208	1.132370e-05	3.767705e-05
GNA13	0.6506869	2.079572e-04	4.633066e-04
ADCY7	0.7754407	3.743024e-08	5.617768e-07
CCR5	0.9650823	2.552337e-05	7.458712e-05
AGT	1.5556811	1.838151e-09	7.794207e-08
EDNRA	2.6090534	6.542054e-14	9.098689e-11
F2R	1.1339267	2.862349e-11	3.831556e-09
GNA15	1.0572752	5.651539e-08	7.341371e-07
GNAI1	1.1647542	4.400881e-08	6.262700e-07
GNAI2	0.5405538	2.467433e-06	1.102025e-05
GNB1	0.5313184	1.851493e-05	5.693667e-05
GNG11	0.5147281	2.739498e-03	4.703045e-03
CXCL8	2.6730278	1.631514e-08	3.235422e-07
ARRB2	0.5230217	8.243569e-04	1.587385e-03
PDE1A	0.9938533	2.524808e-07	2.019652e-06
GNB4	1.1945034	1.232790e-09	5.967161e-08
RGS1	1.5075832	6.736354e-08	8.256981e-07
CCL4	0.7058465	2.345350e-02	3.349445e-02
RGS18	0.7717132	7.811707e-03	1.221190e-02
CXCR4	2.5578410	4.260168e-08	6.137820e-07

There are only three genes with negative fold change values in dataset 16515. Drug can bind these genes with an activator reaction were searched by using a query. As a result of the query, any drug could not be obtained from this group.

Table 6.5. The drugs that are targeting the genes listed in Table 6.2.

	Gene	**Drug**
1	HSP90AA1	TANESPIMYCIN
2	HSP90AA1	ALVESPIMYCIN
3	HSP90AA1	GANETESPIB
4	HSP90AA1	BIIB021
5	HSP90AA1	RETASPIMYCIN
6	HSP90AA1	XL-888
7	HSP90AA1	ONALESPIB
8	OXTR	ATOSIBAN
9	OXTR	RELCOVAPTAN
10	OXTR	RETOSIBAN
11	OXTR	RETASPIMYCIN HYDROCHLORIDE
12	OXTR	BARUSIBAN
13	OXTR	EPELSIBAN
14	OXTR	RO-5028442
15	OXTR	BALOVAPTAN
16	OXTR	CHEMBL306645
17	OXTR	NELIVAPTAN
18	OXTR	CHEMBL296908

Dataset 16515 contains eleven genes with positive fold change values (Table 6.2). In the filtering process for drug selection, the interaction types in the inhibitor group used for these genes. In the drug query, drugs were obtained for two genes: HSP90AA1 and OXTR. As a result of the query, a total of 18 drugs with different names were obtained from different databases.

Table 6.6. The drugs that are targeting the genes listed in Table 6.3.

	Gene	**Drug**
1	S1PR5	SIPONIMOD
2	S1PR5	OZANIMOD
3	S1PR5	FINGOLIMOD
4	S1PR5	ASP-4058
5	S1PR5	ETRASIMOD
6	S1PR5	PONESIMOD
7	UTS2R	UROTENSIN-II
8	UTS2R	CHEMBL192359

There are seven genes with negative fold change values in Dataset 15471 (Table 6.3). In the filtering process for drug selection, the interaction types in the activating group are used for genes. In this group, drugs were obtained for genes with S1PR5 and UTS2R gene symbols. As a result of the query for these two genes, a total of 8 drugs were obtained from different databases.

Table 6.7. The drugs that are targeting the genes listed in Table 6.4.

	Gene	**Drug**
1	CCR5	MARAVIROC
2	CCR5	VICRIVIROC
3	CCR5	LERONLIMAB
4	CCR5	CENICRIVIROC
5	CCR5	CHEMBL207004

6	CCR5	AZD5672
7	CCR5	INCB-9471
8	CCR5	CHEMBL41275
9	CCR5	PF-04634817
10	CCR5	APLAVIROC
11	CCR5	CHEMBL2018969
12	CCR5	ANCRIVIROC
13	CXCR4	PLERIXAFOR
14	CXCR4	MAVORIXAFOR
15	CXCR4	BKT140
16	CXCR4	BURIXAFOR
17	CXCR4	MSX-122
18	CXCR4	POL6326
19	CXCR4	CTCE-9908
20	CXCR4	ULOCUPLUMAB
21	CXCL8	ABX-IL8
22	CXCL8	HUMAX-IL8
23	EDNRA	BOSENTAN
24	EDNRA	AMBRISENTAN
25	EDNRA	SITAXENTAN
26	EDNRA	MACITENTAN
27	EDNRA	ATRASENTAN
28	EDNRA	DARUSENTAN
29	EDNRA	ZIBOTENTAN
30	EDNRA	NEBENTAN
31	EDNRA	CLAZOSENTAN
32	EDNRA	SPARSENTAN

33	EDNRA	AVOSENTAN
34	EDNRA	ENRASENTAN
35	EDNRA	BQ-123
36	EDNRA	TEZOSENTAN
37	EDNRA	APROCITENTAN
38	EDNRA	PD-156707
39	F2R	VORAPAXAR
40	F2R	ATOPAXAR
41	F2R	RIGOSERTIB SODIUM
42	PDE1A	DIPYRIDAMOLE
43	PDE1A	PENTOXIFYLLINE
44	PDE1A	CRISABOROLE
45	PDE1A	VINPOCETINE

Dataset 15471 contains twenty genes own positive fold change values (Table 6.4). In the filtering process for drug selection, the interaction types in the inhibitory group used for genes with positive fold change values were used. In the drug query, drugs were obtained for genes possess, CCR5, CXCR4, CXCL8, EDNRA, F2R, and PDE1A gene symbols. As a result of the query, a total of 45 drugs with different names were obtained from different databases.

CHAPTER SEVEN

CONCLUSION

The aim of the project was to propose new treatments for pancreatic cancer using microarray gene expression profiles for pancreatic cancer-specific normal samples and tumor samples. In addition, owing to the platform created, this study has made it possible for users to research new treatments for other cancer types using cancer-specific microarray gene profiles. Normalization and background correction processes were first applied to the microarray data. Mapping between manufacturer identifiers and gene symbols has been performed. Annotations were added to the new data set after the two data sets were linked. Values with the same annotation values were grouped and averaged. Fold change values and *p*-values were calculated by performing a t-test and created a differentially expressed gene set. The edge list of protein-protein interaction (PPI) network was created with data from the STRING database by choosing high confident interactions. In order to identify the most important genes in this network, personalized Page Rank algorithm was run on the network. The page rank function was executed separately for each data set with the corresponding customized input vectors. The important genes with higher Page Rank values were selected as potential treatment candidates. The rDGBdb library was used for drug selection of drugs that are targeting treatment candidates. Finally, a total of 71 drugs were obtained for treatment candidate of two pancreas cancer microarray data sets. We rearranged the algorithms we applied for pancreatic cancer datasets so that they can be used in different datasets. Then we created a website where users can upload at least two .CEL files, namely cancer samples and healthy controller related to cancer. This interface allows the user to enter their desired p and fold change values. According to these data, appropriate treatments are shown to the user.

The project is innovative in that users can upload microarray gene expression profiles of different cancers to the system and present appropriate treatments to users. This innovative aspect of the project is our advantage over our competitors.

REFERENCES

Masataka Kikuyama, Terumi Kamisawa, Sawako Kuruma, Kazuro Chiba, Shinya Kawaguchi, Shuzo Terada &Tatsunori Satoh. (2018). Early Diagnosis to Improve the Poor Prognosis of Pancreatic Cancer. NCBI

Hyuna Sung PhD,Jacques Ferlay MSc, ME,Rebecca L. Siegel MPH,Mathieu Laversanne MSc et al. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. Volume71, Issue3 Pages 209-249

American Cancer Society. "Key Statistics for Pancreatic Cancer". Access: 18 November 2021. https://www.cancer.org/cancer/pancreatic-cancer/about/key-statistics.html

Prashanth Rawla, Tagore Sunkara & Vinaya Gaduputic. (2019). Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors

Yanfen Ma, Jian Hu, Ning Zhang, Xinran Dong, Ying Li, Bo Yang, Weidong Tian, Xiaoqin Wang. (2016). Prediction of Candidate Drugs for Treating Pancreatic Cancer by Using a Combined Approach. Retrieved: https://doi.org/10.1371/journal.pone.0149896

Jingqi Chen, Ming Ma, Ning Shen, Jianzhong Jeff Xi, and Weidong Tian. (2013). Integration of Cancer Gene Co-expression Network and Metabolic Network To Uncover Potential Cancer Drug Targets.

Yanfen Ma, Jian Hu, Ning Zhang, Xinran Dong, Ying Li, Bo Yang, Weidong Tian, Xiaoqin Wang. (2016). Prediction of Candidate Drugs for Treating Pancreatic Cancer by Using a Combined Approach. Retrieved: https://doi.org/10.1371/journal.pone.0149896

Gregory Glennon. (2000). High-throughput gene expression analysis for drug discovery. Volume 5, Issue 2, 1 February 2000, Pages 59-66. Retrieved: https://doi.org/10.1016/S1359-6446(99)01448-8

Justin Lamb, Emily D. Crawford, Davıd Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael

Reich, Haley Hıeronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A. Clemons, Ru Wei, Steven A. Carr, Erıc S. Lander, Todd R. Golub. (2006). The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. Vol 313, Issue 5795 pp. 1929-1935. DOI: 10.1126/science.1132939 Retrieved: https://www.science.org/doi/10.1126/science.1132939

Hollander M, Wolfe DA, Chicken E (2013) Nonparametric statistical methods: John Wiley & Sons.

Kim JJ, Kurita T, Bulun SE. (2013). Progesterone action in endometrial cancer, endometriosis, uterine fibroids, and breast cancer. Endocrine Reviews, Volume 34, Issue 1, 1 February 2013, Pages 130–162. https://doi.org/10.1210/er.2012-1043 Endocr Rev 34: 130–162. pmid:23303565

Aravind Subramanian , Rajiv Narayan , Steven M Corsello , David D Peck , Ted E Natoli , Xiaodong Lu , Joshua Gould , John F Davis , Andrew A Tubelli , Jacob K Asiedu , David L Lahr , Jodi E Hirschman , Zihan Liu , Melanie Donahue , Bina Julian , Mariya Khan , David Wadden , Ian C Smith , Daniel Lam , Arthur Liberzon , Courtney Toder , Mukta Bagul , Marek Orzechowski , Oana M Enache , Federica Piccioni , Sarah A Johnson , Nicholas J Lyons , Alice H Berger , Alykhan F Shamji , Angela N Brooks , Anita Vrcic , Corey Flynn , Jacqueline Rosains , David Y Takeda , Roger Hu , Desiree Davison , Justin Lamb , Kristin Ardlie , Larson Hogstrom , Peyton Greenside , Nathanael S Gray , Paul A Clemons , Serena Silver , Xiaoyun Wu , Wen-Ning Zhao , Willis Read-Button , Xiaohua Wu , Stephen J Haggarty , Lucienne V Ronco , Jesse S Boehm , Stuart L Schreiber , John G Doench , Joshua A Bittker , David E Root , Bang Wong , Todd R Golub. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Volume 171, Issue 6, 30 November 2017, Pages 1437-1452.e17. Retrieved: 10.1016/j.cell.2017.10.049.

Newman, M. E. J., Girvan, M. (2004). Finding and evaluating community structure in networks. Vol. 69, Iss. 2 — February 2004 . Phys. Rev. E 69 (2), 026113. doi: 10.1103/PhysRevE.69.026113

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., Chen, C. F. (2007). A new method to measure the semantic similarity of GO terms. Bioinformatics, Volume 23, Issue 10, 15 May 2007, Pages 1274–1281, doi: 10.1093/bioinformatics/btm087

Khodade, P., Prabhu, R., Chandra, N., Raha, S., Govindarajan, R. (2007). Parallel implementation of AutoDock. J. Appl. Crystallogr. 40, 598–599. doi: 10.1107/S0021889807011053

Wenying Yan, Xingyi Liu, Yibo Wang, Shuqing Han, Fan Wang, Xin Liu, Fei Xiao and Guang Hu. (2020). Identifying Drug Targets in Pancreatic Ductal Adenocarcinoma Through Machine Learning, Analyzing Biomolecular Networks, and Structural Modeling. Retrieved: https://doi.org/10.3389/fphar.2020.00534

J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K.N. Ross, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease Science, 313 (2006), pp. 1929-1935, 10.1126/science.1132939

A. Subramanian, R. Narayan, S.M. Corsello, D.D. Peck, T.E. Natoli, X. Lu, J. Gould, J.F. Davis, A.A. Tubelli, J.K. Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles Cell, 171 (2017), pp. 1437-1452.e17, 10.1016/j.cell.2017.10.049

Yasaman KalantarMotamedi, Ran Joo Choi, Siang-Boon Koh, Jo L. Bramhall, Tai-Ping Fan, Andreas Bender. (2021). Prediction and identification of synergistic compound combinations against pancreatic cancer cells. Volume 24, Issue 9, 103080, September 24, 2021.

Retrieved: https://doi.org/10.1016/j.isci.2021.103080

Christine A. Iacobuzio-Donahue, Anirban Maitra, Mari Olsen, Anson W.Lowe, N. Tjarda Van Heek, Christophe Rosty, KimWalter et al. Exploration of Global Gene Expression Patterns in Pancreatic Adenocarcinoma Using cDNA Microarrays. (2003) Volume 162, Issue 4, April 2003, Pages 1151-1162

Brianne Petritis, Phd .SAM (Significance Analysis Of Microarray). (2018). Retrieved: https://www.raybiotech.com/learning-center/sam/ [Accessed December 2021]

Layla Abdel-Ilah , Elma Veljovic, Lejla Gurbeta , Almir Badnjevic. Applications of QSAR Study in Drug Design.(2017). Paper ID : IJERTV6IS060241. Volume 06, Issue 06 (June 2017)

LiPeizhen Li, Yueli Tian, Hong Lin Zhai, Lanzhou University, Fangfang Deng. Study on the activity of non-purine xanthine oxidase inhibitor by 3D-QSAR modeling and molecular docking. (2013) 1051:56-65 DOI:10.1016/j.molstruc.2013.07.043

K W Le, J M Briggs. (2001). Comparative molecular field analysis (coMFA) study of epothilones-tubulin depolymerization inhibitors: pharmacophore development using 3D QSAR methods. J Comput Aided Mol Des. 2001 Jan;15(1):41-55. doi: 10.1023/a:1011140723828.

Jingqi Chen, Ming Ma, Ning Shen, Jianzhong Jeff Xi, Weidong Tian. (2013). Integration of Cancer Gene Co-expression Network and Metabolic Network To Uncover Potential Cancer Drug Targets. J. Proteome Res. 2013, 12, 6, 2354–2364

Publication Date:April 16, 2013. DOI: 10.1021/pr400162t

Lin, X.; Yang, F.; Zhou, L.; Yin, P.; Kong, H.; Xing, W.; Lu, X.; Jia, L.; Wang, Q.; Xu, G. A support vectormachine-recursive feature elimination feature selection method based on artificial contrast variables andmutual information. J. Chromatogr. B Anal. Technol. Biomed. Life Sci. 2012,910, 149–155. [CrossRef]

Wikipedia. “Bioconductor”. Last edited: 27 October 2021, at 20:34. Access: 8 December 2021. https://en.wikipedia.org/wiki/Bioconductor

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, e47.

Gordon Smyth. R Documentation. "Introduction to the LIMMA Package". Access: 8 December2021. http://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/limma/html/01Introduction.html

Yu G, Wang L, Han Y and He Q. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 2012, 16(5):284-287.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Mach. Learn. 46 (1-3), 389–422. doi: 10.1023/A:1012487302797

Sunil Ray. "Understanding Support Vector Machine (SVM) algorithm from examples (along with code)". (2017). Access: 09.12.2021. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

Elma Veljovic, Selma Spirtovic-Halilovic, Samija Muratović, Amar Osmanović, Almir Badnjević, Lejla Gurbeta, Berina Tatlić, Zerina Zorlak, Selma Imamović, Đenana Husić, Davorka Završnik. (2017). Artificial Neural Network and Docking Study in Design and Synthesis of Xanthenes as Antimicrobial Agents. DOI:10.1007/978-981-10-4166-2_93. In book: CMBEBIH 2017 (pp.617-626)

Codeacademy."What is Scikit-Learn?". Access: 8 December 2021. https://scikit-learn.org/stable/

KEGG: Kyoto Encyclopedia of Genes and Genomes. "KEGG Overview 1. Genomes to Biological System". Access: 8 December 2021. https://www.genome.jp/kegg/kegg1a.html

Ankit Sahu. C#Corner. "Introduction to Web API". Access: 8 December 2021. https://www.c-sharpcorner.com/article/web-api-in-asp-net/

Anshul_Aggarwal. GeeksforGeeks. “Introduction to Visual Studio”. Access: 8 December 2021. https://www.geeksforgeeks.org/introduction-to-visual-studio/

Yu G, Wang L, Han Y and He Q*. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology. 2012, 16(5):284-287.

Holtzman, Coulter, Vivek, J. D. K. (2021b December 15). Web application requirements - Power Platform. Microsoft Docs.

https://docs.microsoft.com/en-us/power-platform/admin/web-application- requirements

Node.js vs Express.js. (2020, December 9).GeeksforGeeks. Retrieved March 30, 2022, from https://www.geeksforgeeks.org/node-js-vs-express-js/

Clinical trials. (n.d.). World Health Organization. Retrieved March 30, 2022, from https://www.who.int/health-topics/clinical-trials#tab=tab_1

Stitch. (n.d.). Stitch Connect API Reference | Stitch Documentation. Stitch Docs. Retrieved March 30, 2022, from https://www.stitchdata.com/docs/developers/stitch-connect/api#access-token--individual-user

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., & Bork, P. (2008). STITCH: interaction networks of chemicals and proteins. Nucleic acids research, 36(Database issue), D684–D688. https://doi.org/10.1093/nar/gkm795

STRING. (n.d.). API - STRING Help. STRING API. Retrieved March 30, 2022, from https://string-db.org/help/api/

Clinical trials. (n.d.). World Health Organization. Retrieved March 30, 2022, from https://www.who.int/health-topics/clinical-trials#tab=tab_1

HPRD. (n.d.). REST API — INDRA 1.21.0 documentation. REST API. Retrieved March 30, 2022, from https://indra.readthedocs.io/en/latest/rest_api.html#local-installation-and-use

“Download.” Downloads - STRING Functional Protein Association Networks, STRING Database, 2022, https://string-db.org/cgi/download?sessionId=bhrZZYPBLNst.

Thurnherr T, Singer F, Stekhoven DJ and Beerenwinkel N. Genomic variant annotation workflow for clinical applications [version 2; referees: 2 approved]. F1000Research 2016, 5:1963. doi:10.12688/f1000research.9357.2

Wagner AH, Coffman AC, Ainscough BJ, Spies NC, Skidmore ZL, Campbell KM, Krysiak K, Pan D, McMichael JF, Eldred JM, Walker JR, Wilson RK, Mardis ER, Griffith M, Griffith OL. DGIdb 2.0: mining clinically relevant drug-gene interactions. Nucleic Acids Research. 2016 Jan 4;44(D1): D1036-44. doi:10.1093/nar/gkv1165.

“Download.” Downloads - STRING Functional Protein Association Networks, STRING Database, 2022, https://string-db.org/cgi/download?sessionId=bhrZZYPBLNst.

Thurnherr T, Singer F, Stekhoven DJ and Beerenwinkel N. Genomic variant annotation workflow for clinical applications [version 2; referees: 2 approved]. F1000Research 2016, 5:1963. doi:10.12688/f1000research.9357.2

Wagner AH, Coffman AC, Ainscough BJ, Spies NC, Skidmore ZL, Campbell KM, Krysiak K, Pan D, McMichael JF, Eldred JM, Walker JR, Wilson RK, Mardis ER, Griffith M, Griffith OL. DGIdb 2.0: mining clinically relevant drug-gene interactions. Nucleic Acids Research. 2016 Jan 4;44(D1): D1036-44. doi:10.1093/nar/gkv1165.