DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

SURVIVAL TIME PREDICTION OF CANCER PATIENTS

by

Müşerref Ece ERCAN

January, 2018

İZMİR

SURVIVAL TIME PREDICTION OF CANCER PATIENTS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylul University
In Partial Fulfillment of the Requirements for the Degree of Master of Science in Computer Engineering

by

Müşerref Ece ERCAN

January, 2018
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"SURVIVAL PREDICTION OF CANCER PATIENTS"** completed by **MÜŞERREF ECE ERCAN** under supervision of **ASST.PROF.DR. ZERRİN IŞIK** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Zerrin IŞIK

Supervisor

Assoc. Prof. Dr. Derya BİRANT

(Jury Member)

Asst. Prof. Dr. Özge CAN

(Jury Member)

Prof. Dr. Kadriye ERTEKİN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to thank my supervisor Asst. Prof. Dr. Zerrin IŞIK who allowed me to work in this project and I appreciate for her valuable ideas, support and guidance throughout this project.

This work is supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK). Project number is 115C012.

Müşerref Ece ERCAN

SURVIVAL TIME PREDICTION OF CANCER PATIENTS

ABSTRACT

In recent years, in order to reduce noise in experimental data and to add the common role of genes in biological processes into diagnostic and prognostic prediction models, researchers entegrates more than one data type. In this context, many studies have shown that protein interaction networks increase the success of scientific diagnosis. This study aims to find biomarkers that successfully predict the potential survival time of cancer patiens by merging gene transcriptome and protein level data belonging to kidney renal clear cell carcinoma (KIRC) and glioblastoma multiforme (GBM). For this purpose, expression level of mRNA (RNA-seq) and protein (RPPA) data entegrated a with network modelling protein interactions in the human genome. Survival time of patients will be predicted by selecting certain amount of biomarkers and feeding those as inputs to the supervied learning method. For both cancer types, this study showed that our new entegrated method, RPBioNet, outperforms both "only protein" and "only mRNA" methods.

Keywords: Bioinformatics, biomarker, gene expression, RPPA, protein interaction network, survival time prediction, data mining, GBM, KIRC, TCGA

KANSER HASTALARININ YAŞAM SÜRESİ TAHMİNİ

ÖZ

Son yıllarda, deneysel verilerdeki kirliliği en aza indirmek ve genlerin biyolojik süreçteki ortak rollerini tanı-tahmin modeline ekleyebilmek için araştırmacılar birden fazla veri türünü entegre etmektedir. Bu bağlamda son yapılan çalışmalar protein etkileşim ağlarının, bilimsel tanı yöntemlerinin başarısını arttırdığını göstermektedir. Bu çalışmada, böbrek (KIRC) ve beyin (GBM) kanseri hastalarına ait gen transkriptom ve protein seviye bilgilerini protein etkileşim ağları ile bütünleştirerek, hastaların olası yaşam süresini başarıyla tahmin edebilecek belirteçler (biyomarker) bulunması amaçlanmıştır. Bu amaçla transkriptom düzeyindeki mRNA ifadesi (RNA-seq) ve protein (RPPA) verileri insan genomundaki protein etkileşimlerini modelleyen bir ağ modeli ile entegre edilmiştir. Hastaların yaşam süresi, belli bir miktar belirteç seçip, ardından bu belirteçleri gözetimli öğrenme yöntemine girdi olarak vererek tahmin edilmiştir. Geliştirdiğimiz yeni entegre yöntemin, RPBioNet, iki kanser türü için de, "sadece protein" ve "sadece mRNA" yöntemlerinden daha başarılı sonuç verdiği görülmüştür.

Anahtar kelimeler: Biyoinformatik, biyomarker, gen ifadesi, RPPA, protein etkileşim ağı, yaşam süresi tahmini, veri madenciliği, glioblastoma multiforme, böbrek kanseri, TCGA

CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

1.1 Motivation

Recently, advancements in proteomics and genomics have helped us to gather an immense amount of biological data which requires complex computational analysis (Raza, 2012). Bioinformatics, or in other words computational biology, analyse the data related with biomolecules on a very large-scale by inferring structure or generalizations from the data (Luscombe, Greenbaum & Gerstein, 2001). Protein structure prediction, cancer subtype classification using microarray data, gene-expression data clustering, gene classification, protein-protein interaction network classifications can be some examples of that kind of analysis (Raza, 2012).

Cancer patients are cured by oncologists according to the cancer type and stage of the patients. The cancer type and stage are diagnosed by pathologists. Extracting tumor samples with a surgery cannot be possible for some sensitive structures, such as optic nerves or brain stem. Sometimes, elaboration of tumor samples in the lab might not help to predict future health conditions (e.g., total survival time, cancer stage in two years etc.) of a patient. So, there is a need for new diagnostic tests that can predict future health condition of patient by using only patient's blood samples. Such diagnostic tests can also help to design targeted therapy (i.e., personalized medicine) specific to each cancer patient. Therefore, research on personalized medicine is taking more attention in recent years.

If it would be possible to identify information specific to a cancer patient without surgery but with a simple and easy-to-implement blood tests, without decreasing patient's life quality; targeted treatment specific to a cancer patient can be applied. In recent years, studies on this topic are being carried out by many researchers. However, these studies do not one hundred percent accurate, yet.

The vast amount of high-throughput patient data become more accessible in the last decade. The Cancer Genome Atlas (TCGA) Project publishes various patient data for 34 cancer types and regularly enlarges the repository ("TCGA research network" 2017). Researchers can use large patient cohorts for prediction of future health states of patients. In this sense, survival time prediction is quite important topic to develop personalized treatment strategies for patients.

1.2 Problem Definition

The discovery of a certain amount of biomarkers for the prediction of cancer type, subtype, and the probable survival time of cancer patients is a challenging problem for distinct reasons. We suggested several proposals to this problem with computational methods. The essential data source of this research is gene expression and protein level data integrated with protein-protein interaction network that is used as an input for feature selection method. Thus, we will investigate whether integrated data can outperform single data for predicting prognostic biomarkers.

Another significant point is what kind of features we are supposed to use to create a training method. There are mRNA levels of about 2100 genes in a human genome belongs to cancer patients in TGCA database. However, not all these data can be given as inputs to machine learning model. Otherwise, in this much large input space, the patient sample amount for a model to be trained will not be enough. For this reason, the most important "N" features will be selected.

The other issue is to reduce the experimental mistakes caused by gene level measurements and adding causal relationships between transcriptome level genes to the model. Does protein-protein interaction network improve the classification performance? At that point, gene level data will be integrated to protein-protein interaction network. Each node represents a human genome and each path between two vertices represents genomic, physical or functional interaction in a protein-protein network.

Furthermore, another concern is cross validation methods. The most convenient data partitioning method to train the model should be selected.

In a nutshell, in this research we will answer whether we can determine a patient's cancer type and cancer subtype by looking at certain biomarkers without applying surgery to this patient. After prognosis, is it possible to predict the survival time of the patient using the same biomarkers? If it is, what is the success rate of this predication? In which cancer related biological processes are these biomarkers play an active role? All these questions will be investigated with detail.

1.3 Contribution

The ultimate goal of this study is to find a limited number of biomarker genes that can successfully predict the potential survival time of GBM or KIRC patients. For this purpose, gene expression and protein level data were integrated by using PPI (Protein-Protein Interaction) and Personalized PageRank algorithm that provides the highly ranked (i.e., most significant) proteins (Isik & Ercan, 2017). In the next phase, the gene expression data of 20 highly ranked proteins were used to train a supervised prediction model.

The novelty of this study is the integration of PPI networks with gene expression and protein level data to find prognostic biomarkers that can be used for clinical purposes (Isik & Ercan, 2017). Finding only a certain amount of biomarkers rather than many of them will reduce the costs of clinical prognostic tests. Another contrubition is the development of a new feature selection method that combines protein and gene expression data in a random walk method (Isik & Ercan, 2017).

1.4 Organization of the Thesis

This thesis includes 5 chapters and the rest of the thesis is organised as follows:

In Chapter 2, the definition of biomarkers and their relationship with cancer have been detailed. Biological data have been discussed regarding RNA-seq, RPPA. Furthermore, survival time prediction researches on GBM and KIRC cancer typeshave been disgussed.

In Chapter 3, inputs that are given to the algorithms, data processing steps including feature selection, different machine learning methods and cross validation process have been described and network-based functional analysis steps have been explained in detail. Besides, Page Rank and personalized Page Rank algorithms have been discussed.

In Chapter 4, experimental results of the proposed method, RPBioNet, have been illustrated with screenshots and tables which represent the main aspects for the project. Besides, biological results such as best predicted genes and gene ontology analysis results have been shown.

In Chapter 5, the conclusion and future works have been discussed.

CHAPTER TWO
LITERATURE REVIEW

2.1 Cancer

There are two types of tumor: benign and malignant (Bollinger, 2016). Benign tumors are not cancerous and unlike benign tumors, malignant tumors are cancerous. Benign tumors can usually be removed, and they do not metastasize (or spread) to different parts of the body and tissue (Bollinger, 2016; "Malignant and benign brain tumors", 2017). They are mostly localized, and they respond well to therapy. However, malignant tumors are dangerous and usually resistant to therapy (Bollinger, 2016; "Malignant and benign brain tumors", 2017). They infest contiguous tissues and have a tendency to recur after removal and spread to other parts of the body ("Malignant and benign brain tumors", 2017). The differences between benign and malignant tumors are shown in Figure 2.1.

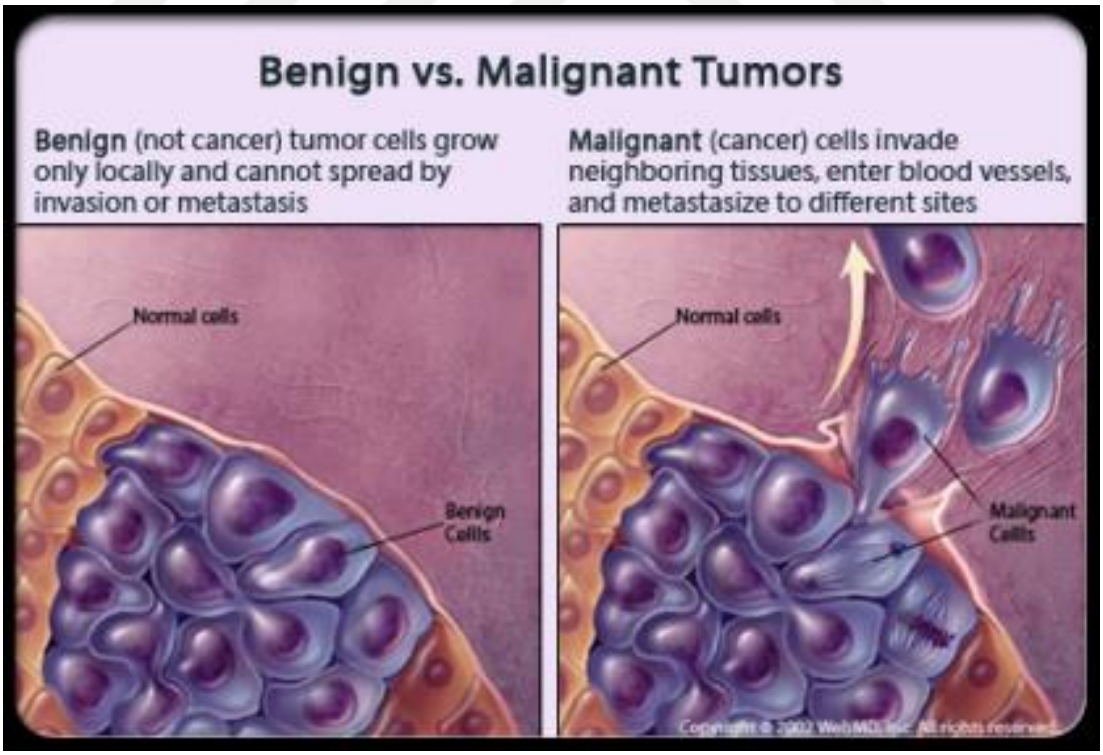

Figure 2.1 Malignant vs benign tumor (Christiansen, 2015)

Malignant tumors are called cancer ("What is cancer?", 2017). Unlike normal cells, cancer cells grow out of control which means they divide uncontrollably ("What is cancer?", 2017). Cancer cells omit signals which order cells to start process called apoptosis (programmed cell death) or to stop dividing ("What is cancer?", 2017).

In multi-cellular organisms like human beings, there must be a homeostatic balance between tissues (Pucci, Kasten & Giordano, 2000). To keep this balance cell proliferation and cell death occurs (Pucci, Kasten & Giordano, 2000). Normally, cells undergo apoptosis when they are damaged or old that makes apoptosis a chance for organisms to get rid of defective or redundant cells (Pucci, Kasten & Giordano, 2000).

Tumor suppressor genes like p53 is associated with apoptosis. Down regulation or mutation of p53 gene brings about lessened apoptosis, thus improved tumor development (Wong, 2011). Mutations arise from either copying DNA damage or failures happened during DNA synthesis(Loeb & Loeb, 2000). This damage is caused by environmental and internal (cellular) origins, which results in genetic instability (Loeb & Loeb, 2000). In tumor cells there are aggregation of multiple mutations (Loeb & Loeb, 2000). There is an unbalance between cell cycle and cell death (or apoptosis) in cancer (Wong, 2011). Inactivation or any mutation in apoptosis represents a major reason for development and advancement of cancer (Kasibhatla & Tseng, 2003). It means that a better understanding of apoptosis can make a contribution to cancer treatment.

Aside from consecutive accretions of genetic mutations, epigenetic mutations are also causes cancer (Virani, Colacino, Kim & Rozek, 2012). Epigenetic change is an alteration in phenotype (genetic traits and environmental aspects) without an alteration in genotype, or in other words DNA sequence (Virani, Colacino, Kim & Rozek, 2012). There is a hot topic in epigenetic called DNA methylation (Virani, Colacino, Kim & Rozek, 2012). Lack of DNA methylation is one of the epigenetic alterations which is characterized in cancer (Virani, Colacino, Kim & Rozek, 2012).

Since these changes are reversible, they give a strong hope for cancer treatment (Virani, Colacino, Kim & Rozek, 2012).

By the National Center for Health Statistics, in 2016, 1.685.210 new cancer occurrence and 595.690 deaths caused by cancer took place in the United States (Siegel, Miller & Jemal, 2016). Cancer mortality rates are declined 23% since 1991 (Siegel, Miller & Jemal, 2016). However, for some cancer types like liver, pancreas, and uterine corpus, fatality rates are rising (Siegel, Miller & Jemal, 2016). Figure 2.2 illustrates top ten most common cancer types forecasted to happen in men and women in 2016:

Estimated New Cases

Males			Females		
Prostate	180,890	21%	Breast	246,660	29%
Lung & bronchus	117,920	14%	Lung & bronchus	106,470	13%
Colon & rectum	70,820	8%	Colon & rectum	63,670	8%
Urinary bladder	58,950	7%	Uterine corpus	60,050	7%
Melanoma of the skin	46,870	6%	Thyroid	49,350	6%
Non-Hodgkin lymphoma	40,170	5%	Non-Hodgkin lymphoma	32,410	4%
Kidney & renal pelvis	39,650	5%	Melanoma of the skin	29,510	3%
Oral cavity & pharynx	34,780	4%	Leukemia	26,050	3%
Leukemia	34,090	4%	Pancreas	25,400	3%
Liver & intrahepatic bile duct	28,410	3%	Kidney & renal pelvis	23,050	3%
All Sites	**841,390**	**100%**	**All Sites**	**843,820**	**100%**

Estimated Deaths

Males			Females		
Lung & bronchus	85,920	27%	Lung & bronchus	72,160	26%
Prostate	26,120	8%	Breast	40,450	14%
Colon & rectum	26,020	8%	Colon & rectum	23,170	8%
Pancreas	21,450	7%	Pancreas	20,330	7%
Liver & intrahepatic bile duct	18,280	6%	Ovary	14,240	5%
Leukemia	14,130	4%	Uterine corpus	10,470	4%
Esophagus	12,720	4%	Leukemia	10,270	4%
Urinary bladder	11,820	4%	Liver & intrahepatic bile duct	8,890	3%
Non-Hodgkin lymphoma	11,520	4%	Non-Hodgkin lymphoma	8,630	3%
Brain & other nervous system	9,440	3%	Brain & other nervous system	6,610	2%
All Sites	**314,290**	**100%**	**All Sites**	**281,400**	**100%**

Figure 2.2 Top ten cancer types for the Estimated New Cancer Cases and Estimated Deaths according to sex in the US in 2016 (Siegel, Miller & Jemal, 2016)

2.1.1 Biomarkers and Their Relationship with Cancer

Biomarkers (i.e., biological markers), are measurable and objective signs of a biological state or condition (Strimbu & Tavel, 2010). Biomarkers are clinically measured, and it can be a signal change in the state of a protein that correlates with the development or risk of an illness, or sensitivity of the illness to an applied treatment (Guenther, Hauser, & Huss, 2016). Biomarkers can be biological features or particles that may be discovered or measured in any parts of the body such as the tissue or blood (Mayeux, 2004).

For example, body temperature is a widely known biomarker for fever or blood pressure is used to detect the possibility of stroke (Kropotov, 2016). High-sensitivity C-reactive protein (HSCRP) is an inflammatory marker that is used for detecting atherothrombosis of the cerebral and coronary vessels (Ridker & Silvertown, 2008).

The main characteristics of an optimal biomarker are listed below:

- Safely obtained and easy to measure (Kufe et al., 2003).
- Cost efficient and high throughput to follow up (Kufe et al., 2003).
- High accuracy and disease specificity (Kufe et al., 2003).
- High accuracy and disease specificity which results in a low false-positive rate (FPR) and false-negative rate (FNR) (Kufe et al., 2003).
- Consistent across gender (Ananya, 2014).

Biomarkers play a key role in the monitoring, follow-up, diagnosis and monitoring of many cancer types (Kobayashi et al., 2012). The aim of anticancer therapies is targeting and killing cancerous cells while sparing healthy cells (Delalat et al., 2015). The discovery of a biomarker can help oncologists to apply treatments targeted only specific cancer cells which reduces the risk of harm and cost of the treatment (Ananya, 2014).

Genetic mutations (i.e., genetic abnormalities), are the primary cause of cancer (Kurzawski et al., 2012). Hence, predefined RNA or DNA biomarkers may most likely help oncologists to discover and treat specific types of cancers (Kurzawski et al., 2012). If a genetic test displays that presence or levels of a certain biomarker is disparate than what is found in healthy tissues, it might show that the cancer is contingent on the change in that biomarker ("Biomarker Testing" 2017). For instance, p53 gene is very well-known tumor suppressor gene which is often mutated in cancer cells (Hong, van den Heuvel, V Prabhu, Zhang & S El-Deiry, 2014). By looking at some researches,p53 gene is a significant biomarker of survival in osteosarcoma, which is a malignant bone tumor with low survival rates (Fu et al., 2013).

2.2 Biological Data

2.2.1 RNA-Seq (RNA sequencing)

Information such as personal traits, behaviors of each single cells, the color of a person's hair, the scent of a violet, the way in which bacteria infect a skin cell, etc are encoded by DNA (Finotello & Di Camillo, 2015). Using this information, cells are able to gain and convert (or translate) particular instructions through gene expression by switching set of genes on and off (Finotello & Di Camillo, 2015). This encrypted info in the selected genes transcribed (or copied) into messenger RNA(mRNA), which can be translated into proteins (Finotello & Di Camillo, 2015). This means that the group of RNAs copied in a certain time and circumstance tells the present state of a cell and able to report pathological structure of disease (Finotello & Di Camillo, 2015).

RNA-sequencing is a RNA profiling technique which makes it possible to measure and assess the similarities and differences between gene expression patterns at high resolution (Finotello & Di Camillo, 2015). It provides immense amount of data for transcriptomics researches (Finotello & Di Camillo, 2015). A gene expression is illustrated in Figure 2.3.

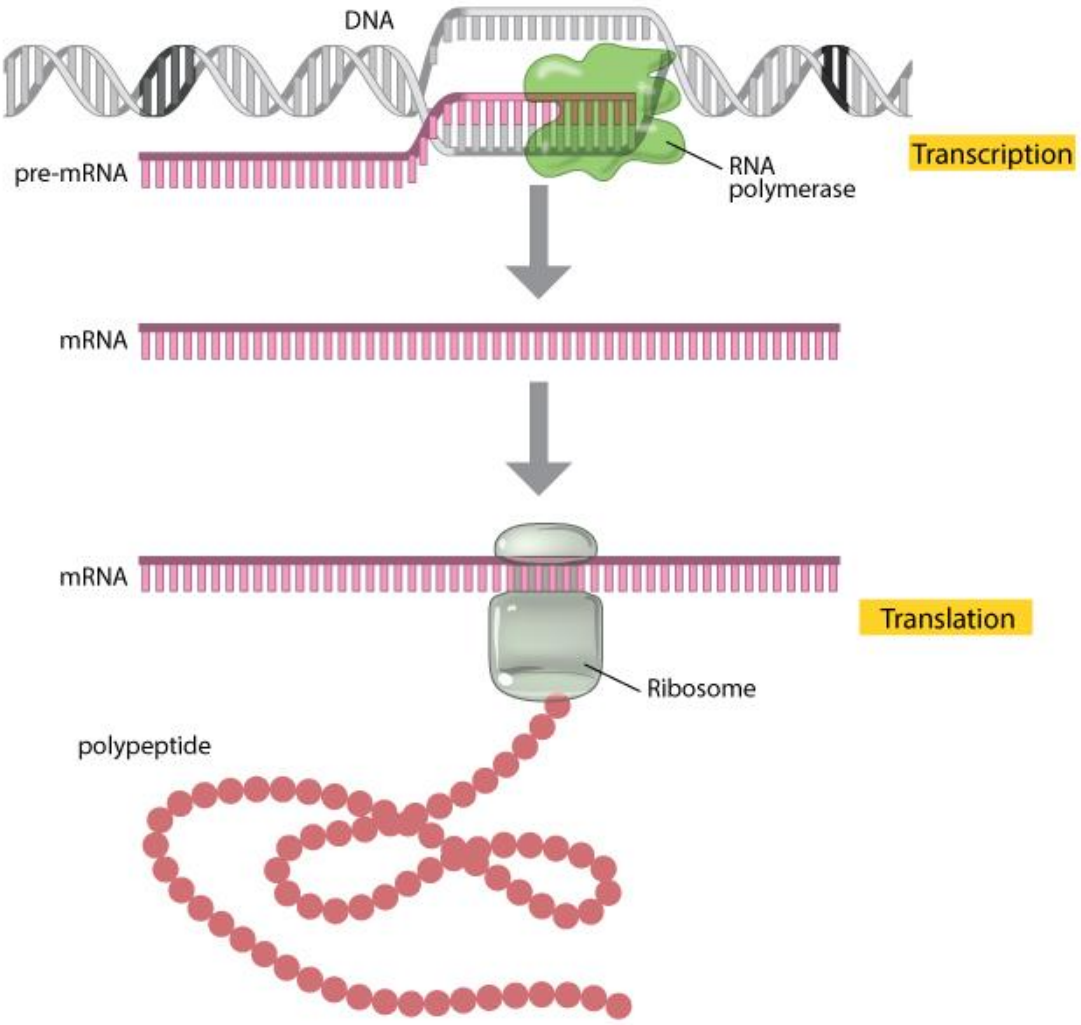

Figure 2.3 mRNA becomes a single-stranded replica of the gene. As the next step, it will be translated into a protein (Clancy & Brown, 2008)

There are some significant stages in RNA-Seq data analysis such as experimental analysis and design, sequence quality analysis, alignment reading, measuring gene expression, RNA-Seq data visualization, differential gene expression, alternative splicing, functional analysis of gene lists (Conesa et al., 2016). Figure 2.4 shows three significant data analysis processes for both microarray and RNA-Seq technologies.

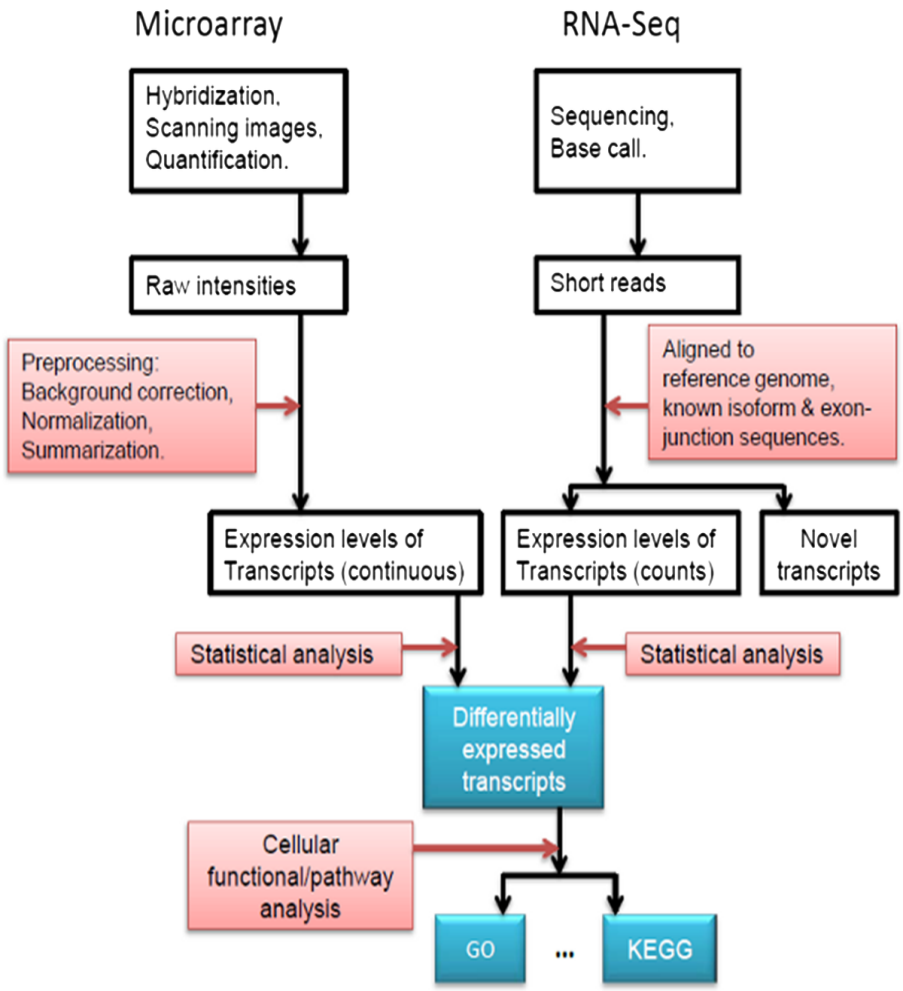

Figure 2.4 A brief summary of microarray and RNA-Seq technologies workflow (Corney, 2013)

A recent study by Zhao et al. (2014) showed that even if there is a high correlation between the RNA-Seq and microarray technology, RNA-Seq outperforms microarray technology in terms of transcriptome profiling. RNA-Seq performs better at discriminating extremely important isoforms, labeling genetic variants, and discovering more differentially expressed genes.

2.2.2 Reverse Phase Protein Arrays (RPPA)

The reverse phase protein array (RPPA) supplies gene-expression data for a predetermined group of proteins, through a group of cell line or tissue samples (O'Mahony et al., 2013). RPPA shows a way to get the state of signal transduction

(the transfer of genetic material from one cell to another) pathways in either normal or diseased cells (Creighton & Huang, 2015). RPPA data may be integrated with other molecular profiling platforms to discover more complete molecular analysis of the cell (Creighton & Huang, 2015).

The RPPA technology is a kind of protein microarray that is dedicated from gene expression microarrays, which is obtained by printing DNA molecules on microscope slide, and immunoassays, which is for discovering protein expression by antigen and antibody interplay (Creighton & Huang, 2015).

In short, RPPA is an effective antibody-based, cost effective, quantitative way for targeted proteomics which can be applied to large datasets (Zhang, Chen, Huang, Zhang, Kong & Cai, 2015). Since protein expression is more reliable than gene-expression, Zhang et al. (2015) have picked RPPA data to classify cancer subtypes and to find a small number of key genes to discriminate different cancers. They proposed a method to categorize the patient samples into ten cancer types on the basis of the RPPA data with help of the sequential minimal optimization method. First, they applied feature selection to pick 23 significant proteins out of 187 proteins applying minimum redundancy maximum relevance feature selection and incremental feature selection methods on the training dataset. Their workflow can be seen in Figure 2.5.

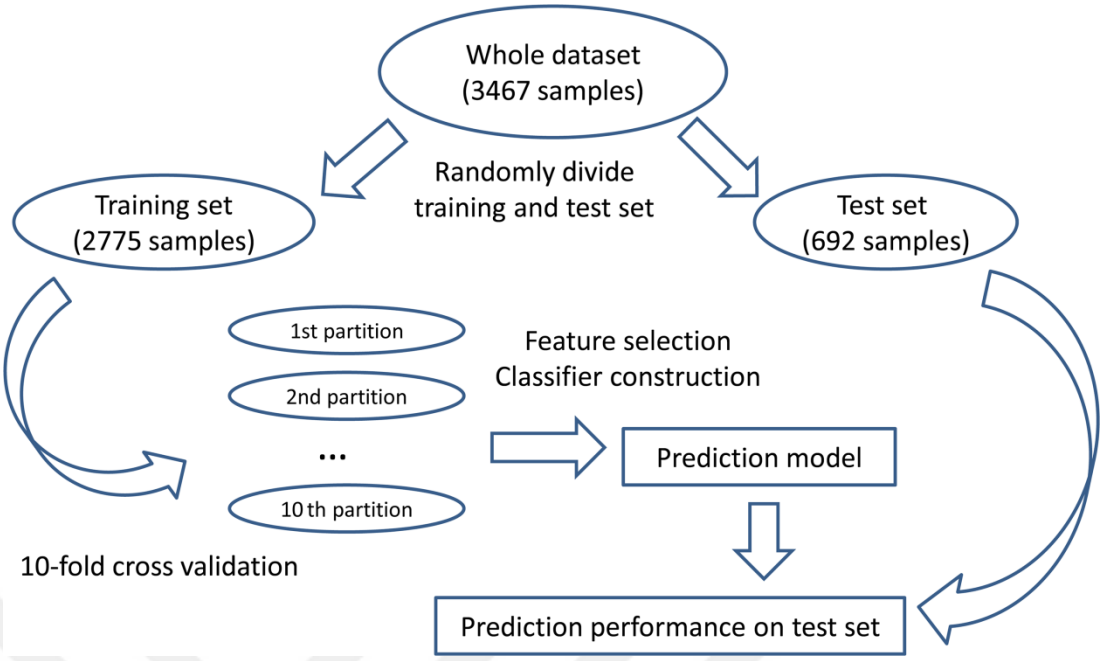

Figure 2.5 Model development and assessment. By dividing data set into a train and a test set and then applying 10-fold cross-validation on 10 partitioned training set, the features are picked and prediction model is built. (Zhang, Chen, Huang, Zhang, Kong & Cai, 2015)

Akbaniet al. (2014) mixed the proteomic data with transcriptomic and genomic data to label common attributes, dissimilarities, signaling pathways, biological network of tumor origins by using TGCA "pan-cancer" patient samples. They also improved biomarker detection by decreasing tissue-specific signals. This combining method is called “replicates-based normalization” (RBN), which is a novel approach for deciding diagnostic, curative importance of proteome (Akbaniet al., 2014).

RPPA workflow is shown in Figure 2.6. RPPAs can be used for personalized medicine (Malinowsky, Wolff, Gündisch, Berg & Becker, 2011). By extracting proteins from formalin-fixed, paraffin-embedded (FFPE) tissues, RPPA can provide ways to measure curative targets and prognostic biomarkers in the future (Malinowsky, Wolff, Gündisch, Berg & Becker, 2011).

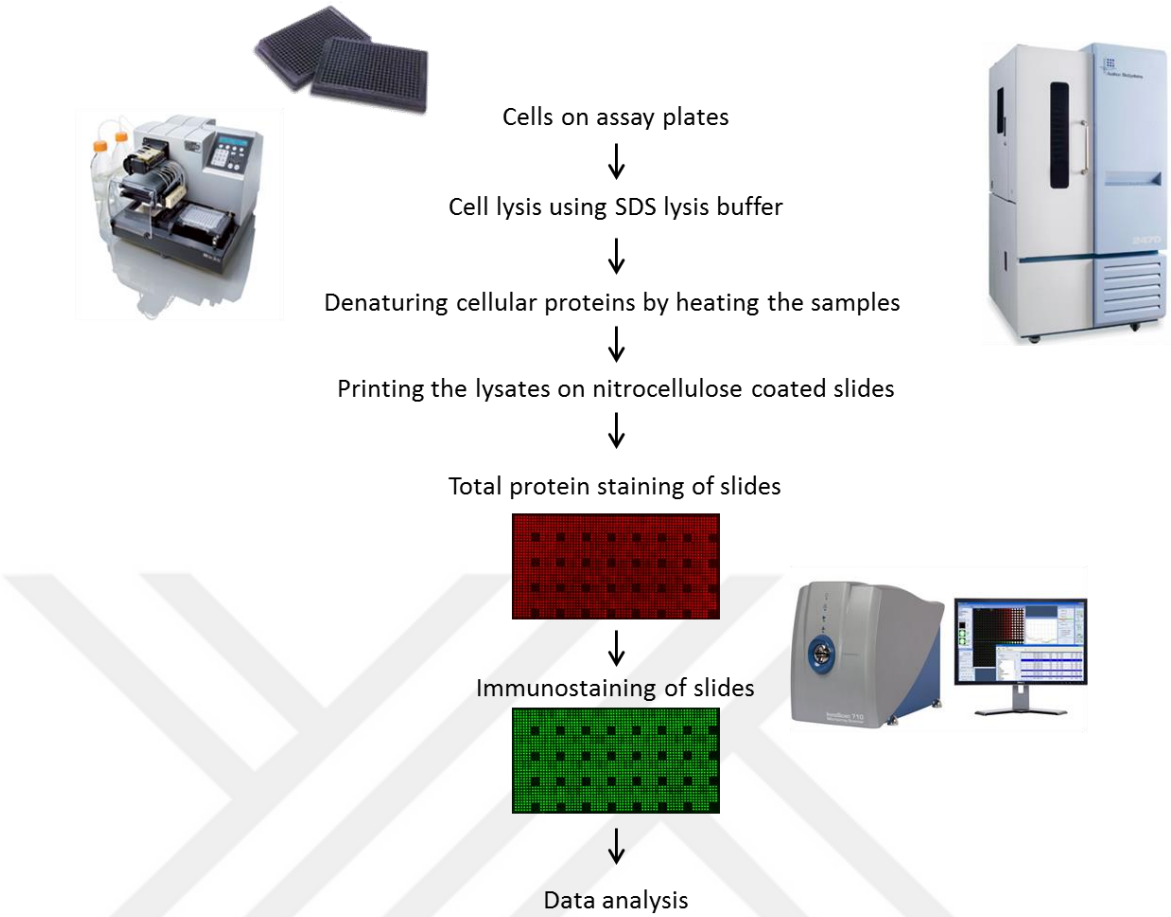

Figure 2.6 Workflow of RPPA ("Example of a RPPA workflow", 2017)

2.3 Survival Time Prediction

There exists FDA approved and commercial diagnostic kits such as “Mammaprint” (Beumer, Witteveen, Delahaye, Wehkamp, Snel, Dreezen & Linn, 2016), “Oncotype Dx” (Buus et al., 2016), and “Prosigna Breast Cancer Prognostic Gene Signature Assay” (Nielsen et al., 2014). These genomic tests analyse a sample of a malignant tumour and identify the activity level of certain genes to predict the recurrence rate or the risk of metastasis in breast cancer. Although such kits are reliable for the clinical usage, they are not 100% accurate yet.

Glioblastoma multiforme (GBM) is the most common and aggressive malignant brain tumour (i.e., gliomas) (Bowman & Joyce, 2014). Glioma is a form of tumour

which develops in the brain and spines ("Glioma", 2017). There are four grades of gliomas, and grade 4, glioblastoma multiforme, is the most aggressive one (Holland, 2000). Region of tumour cells within brain varies, hence the name is multiforme (Holland, 2000). This makes the treatment of GBM even more difficult, it cannot be treated by surgery (Holland, 2000). After diagnosis, total survival time of GBM patient is less than a year. The National Cancer Institute showed that 22.850 adults (12.630 men and 10.280 women) were diagnosed with brain and other nervous system cancer in 2015, and unfortunately 15.320 of these patients died in the same year ("Glioblastoma Multiforme" 2017). Figure 2.7 illustrates MRI scans of a GBM with recurrence in 6 months after surgery (Holland, 2000).

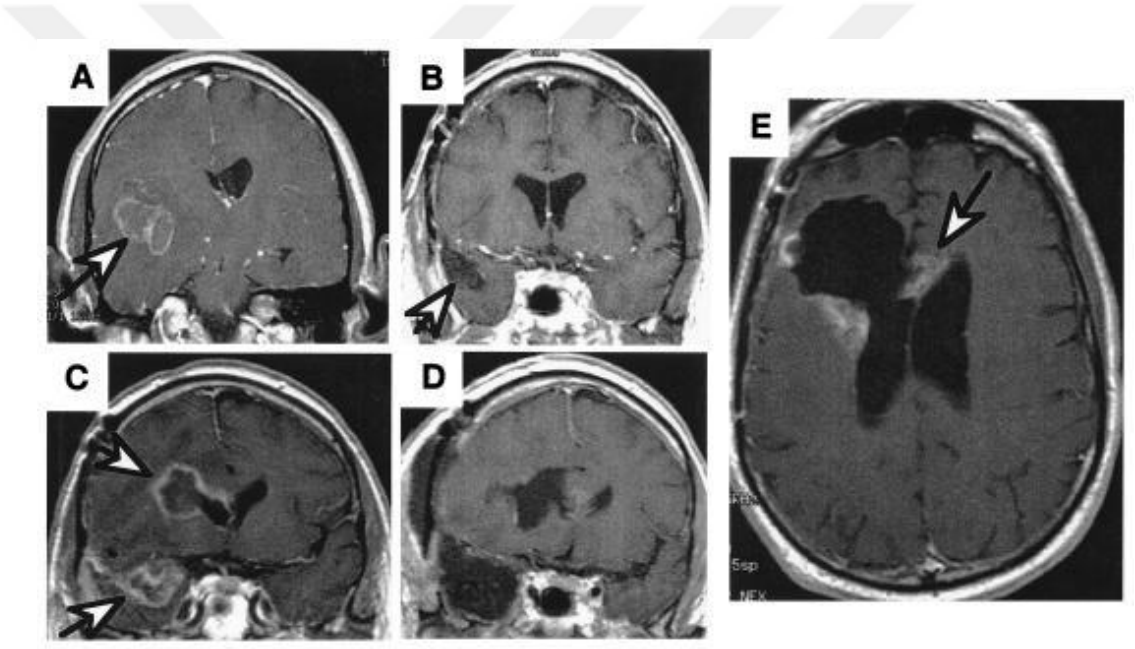

Figure 2.7 MRI scans of a patient with recurrent GBM. Picture (A) shows scan before surgery, (B) shows scan after radiotheraphy and surgery, (C) shows recurrence of GBM after surgery in 6 months, (D) shows recurrence of tumors after cutting them out, (E) shows scan after 3 months, tumor completely spreads out (Holland, 2000)

The occurrence rate and fatality of kidney cancers have been increasing all around the world (Edwards et al., 2014). Kidney renal clear cell carcinoma (KIRC) is the most common type (90–95% of cases) of kidney cancers (Kush, 2014). KIRC is a tumour where VHL gene, which is rarely mutated in other types of tumours, is often inactive (Brugarolas, 2014) in this type of tumour. It is described by a lack of early warning signs, and resistant to chemo and radiation therapies (Kush, 2014).

Zhan et al. (2015) analysed RNA-expression data of KIRC patients for discovering the detailed connection between gene-expression level and diagnosis of these patients. They claimed by carrying out Cox regression and Kaplan-Meier analysis, a five-gene signature could help to predict the survival time (Zhan et al., 2015).

Bie et al. (2011) showed that RNA levels of spindle assembly checkpoint (SAC) genes (BUB1, BUB1B, BUB3, CENPE, MAD1L1, MAD2L1, CDC20, TTK) are related with the grade of glioma and six of them highly related with survival time of GBM patients.

Shen et al. (2016) have introduced a statistical method called SURVIV (Survival analysis of mRNA Isoform Variation), which outruns Cox regression survival analysis, to determine mRNA isoform variation related to cancer patient's (including KIRC and GBM) survival time.

Alexiou et al. (2014) have shown that neutrophil-to-lymphocyte ratio (NLR) has common availability and low-cost. Therefore, neutrophil-to-lymphocyte ratio can be used as a biomarker of GBM aggression and diagnosis.

Many studies have shown that protein-protein interaction (PPI) networks increase the success of cancer diagnosis (Leiserson et al., 2015; Pe'er & Hacohen, 2011; Safari-Alighiarloo, Taghizadeh, Rezaei-Tavirani, Goliaei & Peyvandi, 2014).

Li et al. (2012) have suggested that genes identified from both gene expression profiles, which constructed from protein-protein interaction data, and the shortest path analysis of weighted functional protein association network have found more cancer related genes compared to the classic gene expression analysis.

Smedley et al. (2008) have represented a method for prioritization of candidate disease genes with the help of a global distance measure based on random walk with

restart that characterizes the similarity between genes in protein-protein interaction networks and then proposes new candidates based upon this similarity to known diseases genes.

Study by Zhang et al. (2016) showed that mRNA expression and DNA methylation features provided the most contribution for patient survival, followed by CNV and miRNA features. By determining the importance of the contribution of genes to patient survival considering n-layered regulatory mechanisms such as CNV, DNA methylation, mRNA and miRNA expression and analysing sub networks of the genes related to survival in protein-protein interaction network; a n-dimensional sub network landmark for cancer by combining cancer genomics and interactome data, which illustrated as protein-protein interaction network nodes perturbed by multiple genetic and epigenetic events related to patient survival, is constructed (Zhang et al., 2016).

In breast cancer tissue, Ren et al. (2016) suggested that up-regulated genes related with cell cycle and extracellular matrix interaction causes abnormal breast cancer cycle along with cancer metastasis by constructing protein-protein interaction network encoded by the differentially expressed genes (DEGs) to create the signal transduction network and transcriptional regulatory network. The transcription factor and its multiple downstream regulators, which remarkably higher express in cancer tissue, are the key factor in growth of breast cancer (Ren, Li, Wu, Feng & Li, 2016).

CHAPTER THREE
METHOD

3.1 System Overview

A novel classification method that intent to identify a limited number of protein biomarkers which predict the possible survival time of a cancer patient in a successful manner is established. Overall the new method, which is called RPBioNet, has four main stages (Figure 3.1). In the first stage, patients' samples are split into two parts randomly: 70 percent as training set and remaining 30 percent as test set. For each selected training set, biomarker selection process and supervised learning method is applied. Secondly, the random walk-based Personalized PageRank algorithm runs a human protein-protein interaction network to the most defining N biomarkers (features) whose gene expression values are later used to train a machine learning model. As a third step, the machine learning model is trained by Support Vector Machine (SVM) and Random Forest (RF) supervised machine learning methods to predict new patients' cancer subclass (long-, or short-term). In the end, estimation ability of each trained model is tested on a test data set which is not used when training the model before and accuracy percentage is calculated.

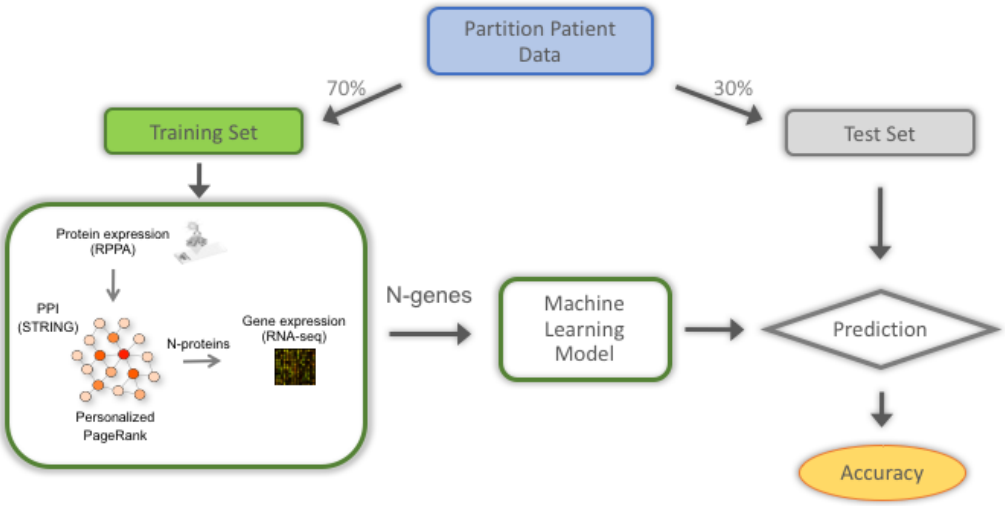

Figure 3.1 The workflow of RPBioNet. The personalized PageRank algorithm is applied on a protein-protein interaction network to uncover the most predictive proteins in the RPPA data. Later, the mRNA data of the 20 selected features are used to train machine learning methods (Support Vector Machine, Random Forest). The performance is calculated over the unseen data and the accuracy is computed. This scheme is repeated 500 times (Monte Carlo cross-validation) and the average accuracy of all iterations is reported as the overall performance of the method

3.2 Data Pre-processing

The data used in this research study were gathered from TGCA ("TCGA research network" 2017) database and saved to a local server. Gene transcriptome (RNA-sequencing) and protein level (Reverse Phase Protein Array-RPPA) data for glioblastoma (GBM) and kidney renal clear cell carcinoma (KIRC) patients were downloaded to a local server via “TCGA Assembler” library in the R-Bioconductor on May 2015. R programming language is used for coding the whole system. As an IDE, RStudio is used. Noise removal and normalization were applied to use these data as an input to supervised learning method. The missing gene and protein data were removed. Since normalization is highly important for clustering of datasets, rows and columns are normalized. Then z-score transformation is applied to gene expression and protein level data. Thus, the raw data for two different cancer types is

converted to z-scores. Those steps will be explained in detail in the following paragraphs.

The RNA-seq data is downloaded by means of the "DownloadRNASeqData" function (parameter called "assayPlatform" was set to "RNASeqV2" and the "dataType" parameter was set to "rsem.genes.normalized.results"). Later, the "ProcessRNASeqData" function is used to get the gene symbol and to extract the normalized count value of each gene. The normalized gene counts are converted into the logarithmic level with the help of the "log2" function. The RPPA data is downloaded to a local server via the "DownloadRPPAData" function with default parameters. After then, the "ProcessRPPADataWithGeneAnnotation" function is applied to obtain gene symbols and protein antibody names. After then, the "ProcessRPPADataWithGeneAnnotation" function is applied to obtain gene symbols and protein antibody names. If the expression of a protein or gene could not be calculated for patients, such proteins or genes were deleted from the data matrix. Later, the patients with survival time, RNA-seq and RPPA data available were preserved during the rest of the analysis. This filtering method significantly lowered the total number of patient samples (KIRC patients were lowered from 417 to 243, GBM patients were lowered from 174 to 35). To normalize RPPA and RNA-seq data, each protein/mRNA level is centralized around a zero mean by subtracting the mean value of patients' samples from the distinctive protein/mRNA level sample. Next, each patient sample is normalized by applying the z-score formula:

$$z.score(i,j) = \frac{x_i - \mu}{\sigma} \tag{3.1}$$

where x_i is the expression value of a patient i, μ represents the population mean, and σ represents the standard deviation of protein/mRNA samples for the patient i. The *z.score(i,j)* demonstrates the normalized expression value of the patient i for a gene j. The z-score transformation is applied for both mRNA and protein level data.

Table 3.1 shows the amount of patients in GBM and KIRC datasets. 24 out of 35 GBM patients are labelled as long-term survivors (LTS) (i.e., can live over 3 years)

whereas 11 of them are labelled as short-term survivors (STS) (less than a year) (Adeberg, Bostel, König, Welzel, Debus & Combs, 2014). 90 out of 243 patients with KIRC patients are labelled as long-term survivors (LTS ≥ 5 years) and 153 of them are labelled as short-term (STS < 2 years) (Choueiri et al., 2007).

Table 3.1 Distribution of data used in this study. The aim is to predict the survival time of KIRC and GBM patients, so two different subclasses are: long-term and short-term survival

Cancer type	**Patient Class**	**Total # of patient class**	**Total # of patients**	**Total # of genes**	**Total # of Proteins**
GBM	Survival time (long-term, short-term)	24 (long-term), 11 (short-term)	35	19080 (gene)	183 (protein)
KIRC	Survival time (long-term, short-term)	90 (long-term), 153 (short-term)	243	20189 (gene)	166 (protein)

Protein level and mRNA gene expression data were provided as the inputs to machine learning algorithms. In order to reduce experimental flaws and integrate causal relationships between the genes into classifier models, a protein-protein interaction (PPI) network was used as a secondary data source. For this purpose, the functional protein interaction information - from the STRING ("STRING: functional protein association networks", 2017) database - was downloaded to a local server, and after that pre-processing and filtering were applied to the data. Initially, only human proteins and corresponding interactions were chosen. Next, interactions with a confidence score 900 or above were preserved in the network. The human PPI network was consisted of roughly 200.000 edges (interactions) and 10.600 nodes (proteins).

3.3 Feature Selection

It is intended to integrate miscellaneous data sources (for instance RPPA, RNA-seq, PPI network) for much better classification model for survival time prediction.

For that, a certain amount of biomarkers were derived from a human PPI network which was later fed to the machine learning method as the input. The most meaningful biomarkers (for example, features) should be extracted out of thousands of genes; or else in a huge feature space, the amount of patient samples would not be adequate to effectively train the model. Furthermore, selecting only a certain amount of biomarkers will cut down the costs of potential prognostic tests in future. For that reason, the first step in creating new method was the selecting of *N* features that were chosen by running a random walk on the PPI network, in which each vertex portrays a gene/protein and an edge between two vertices portray physical, functional or genomic interactions among them. RPBioNet applies the Personalized PageRank algorithm which is the random walk adaption. The PageRank algorithm was originally developed by Google as a web search engine that imitates the behaviours of a random surfer on the Internet (Page, Brin, Motwani & Winograd, 1999).

In the Personalized PageRank, the random web page selection relies on a provided probability distribution. In this study, the rank (in other terms, importance) of a vertex in the PPI is determined by the following equation:

$$p = (1 - d).p.K + d.r \tag{3.2}$$

where *p* gives the rank value of a vertex after the running of the algorithm is ended, *d* means the damping factor which is the probability of continuing to visit other vertices in the network throughout the random walk, *K* represents the adjacency matrix (or in other terms, edges) of the PPI network, and finally *r* shows the vertex selection probabilities during making random choices in the walk. In this research, protein level information was used to construct the *r* vector of the Personalized PageRank algorithm. A t-test is applied to transform z-score of a protein *i*, which is sampled over entire patients' data in the training set, into a probability value of the protein *i*. For this reason, z-score values of proteins are provided to the t-test as the input to calculate the considerable difference between the short and long survival classes. Afterwards, the t-statistic value of the protein *i* was mapped to a value between 0 and 1, which was selected as the initial probability value of the node *i*. If

protein level data of any vertex are not supplied, the initial probability value of this vertex is assigned to zero. In biological terms, proteins with higher protein levels will be selected with higher probabilities while algorithm making random choices. The rank value *p* displays the importance of each protein in the PPI network. The proteins that have higher rank values are chosen as the most significant *N* features (or in other terms biomarkers). The impact of post-transcriptional modifications in the cell may be more precisely measured via protein data; so, protein-level measurements in the *r* vector are used by the feature selection algorithm. Subsequently, z-scores (the gene expression values) of the selected *N*-proteins were supplied as the input of the machine learning method. The "page_rank" function in the "igraph" package of R-Bioconductor was adopted to compute the rank values of proteins in PPI network.

3.4 Machine Learning Algorithms

The learning model is trained with *N*-biomarkers that were chosen by the Personalized PagePank algorithm. The aim is to foresee the most probable survival class (short- or long-term) of a new cancer patient. Two of recent and common algorithms: random forest (RF) and support vector machines (SVM) are preferred to use as machine learning methods.

3.5 Support Vector Machines (SVMs)

SVM is a supervised machine learning technique which is helpful for solving regression and classification problems. Fundamentally, it calculates the optimum separating hyper-plane, that maximizes the margin between the samples of two classes provided by the training set. That way we are provided more improved classifier model, which is less influenced by outliers. The prediction accuracy of it is mostly high, keeps working when training data contain some flaws, and has fewer over-fitting problem compared to the other methods (Statnikov, Wang & Aliferis, 2008), SVM is selected as one of the machine learning methods. SVM is able to work with both nonlinear and linear datasets. Support vector machine is a kernel-based algorithm (Amari & Wu, 1999). A kernel function transforms the input data to

a high-dimensional problem space. It can be linear or nonlinear. Besides a kernel function is a similarity function that we provide to a machine learning method (Amari & Wu, 1999). It requires two inputs and reveals how similar they are. Kernel functions are used to transform the input space into a high-dimensional space to supply a better separation between class samples (Figure 3.2). With the use of kernel functions, the algorithm can operate in a higher-dimensional space without explicitly mapping the input points into the space. This is quite beneficial, since occasionally our higher-dimensional feature space can be infinite-dimensional and therefore unfeasible to compute. Selecting the right kernel depends on what sort of information/feature we are hoping to extract about the data (Souza, 2010). For instance, a polynomial kernel models feature combinations up to the order of the polynomial (Souza, 2010). In this research Radial Basis Function (B-Spline) kernel which selects hyperspheres (or circles) - compared to linear kernel which lets select only hyperplanes is used.

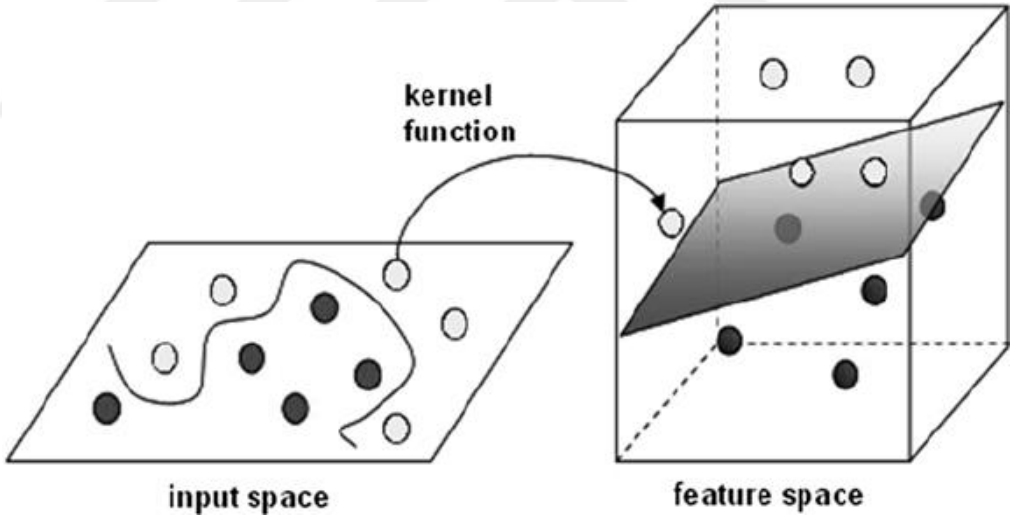

Figure 3.2 The kernel function maps the two dimentional non-linear feature space into a three dimensional feature space. This way, the training set becomes linearly separapable (Schultebraucks, 2017)

When the data is more complicated divided, using just a liner classification method is not enough. Luckily, SVMs can do both linear and non-linear classification based on the linearity of a dataset. More details about these two

classification types will be given in the following sections. Advantages of SVMs can be listed as follows:

- SVMs can ignore outliers and works with both linear and non-linear data
- Most of the time, prediction accuracy is high
- Robust, it can keep working even if training set contains flaws
- In case where the number of dimensions is greater than the number of examples, SVMs are efficient
- Compared to other methods SVMs have less overfitting problem (Ray, Jain, Blog & Saraswat, 2016).

Disadvantages of SVMs can be listed as follows:

- When the data set is wide, training time takes long
- Memory-intensive
- SVMs cannot directly provide statistical results, to compute statistical outcomes we need cross-validation techniques(Ray, Jain, Blog & Saraswat, 2016).

In R-Bioconductor, the package "e1071" has an interface to "libsvm" package. The "svm" function is applied to train a SVM with training datasets. The "c" parameter in "svm" function illustrates a level to prevent misclassification of each training sample. "c-classification" with the "Radial Basis Function (RBF) kernel" is used to stay away from problems caused by tuning. To determine prediction performance of the SVM, the "predict" function is used with test samples. In this research a new feature selection strategy is introduced, so any optimization for the SVM parameters is not made intentionally.

3.5.1 Linear SVM

When segregating two classes, there can be multiple lines/hyper-planes that differentiates the two classes successfully. The intention of SVM is to maximize the

margin/distance from hyperplane to the nearest data point of either class (Figure 3.3). The nearest data points that define the hyperplane are called support vectors.

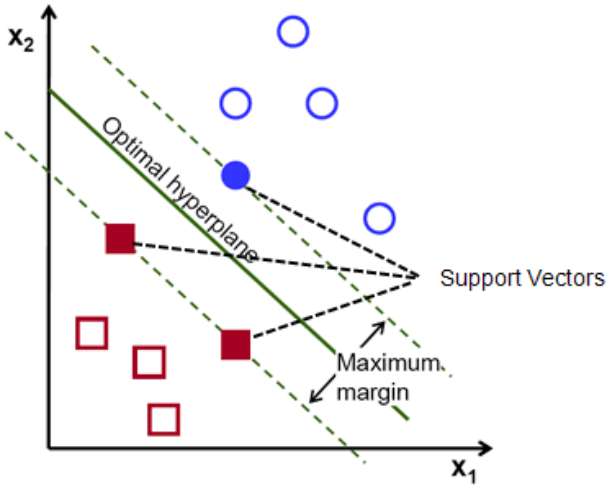

Figure 3.3 Optimal hyperplane between two different classes (Ray, Jain, Blog & Saraswat, 2016)

For instance, in Figure 3.4, hyperplane *C* gives the highest margin compared to hyperplane *A* and *B*. Another reason for choosing the hyperplane with the highest margin is robustness (Ray, Jain, Blog & Saraswat, 2016). If we choose a hyperplane with low margin then there is high risk of misclassification (Ray, Jain, Blog & Saraswat, 2016).

If a separating hyperplane has maximum distance from data points of either classes, like all the data points in each side of the hyperplane have to be of the same class this is called hard margin. This assumes that data is not noisy, and we can find a perfect classifier which will have zero error on training set. However, data may have some errors and we may willing to ignore them to get a better solution. To overcome this problem, we allow for some misclassifications and some data points to be on the wrong side of the hyperplane, so the training error will not be zero, but the average error over all points is minimized. This is called soft margin.

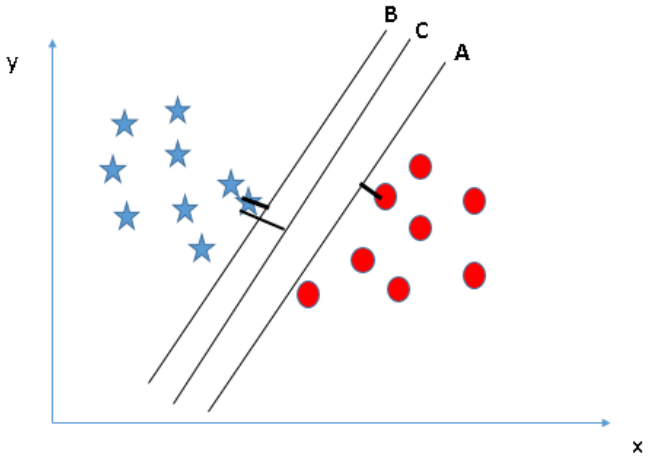

Figure 3.4 Hyperplane *C* is the optimal hyperplane (Ray, Jain, Blog & Saraswat, 2016)

Assume that we have linearly separable data. We pick two parallel hyperplanes, which segregates the two classes, such that distance between two lines is maximum (Figure 3.5).

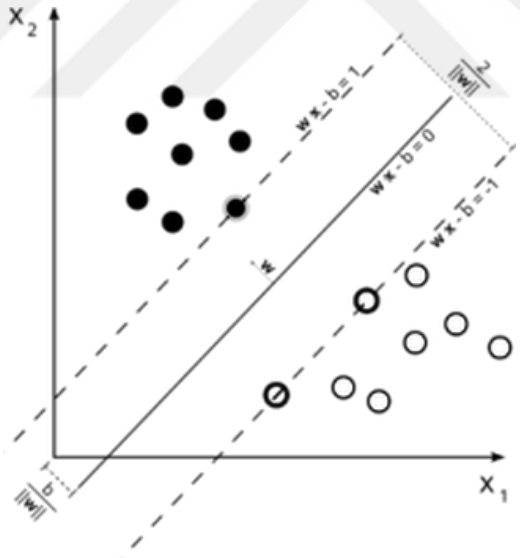

Figure 3.5 Maximum margin hyperplane (Saxena, 2017)

$$\vec{w}.x_i - b \geq 1 \; if \; \theta_i = 1 \tag{3.3}$$

$$\vec{w}.x_i - b \leq 1 \; if \; \theta_i = -1 \tag{3.4}$$

where $||\vec{w}||$ is a vector to hyperplane, θ_i symbolizesclasses and x_i symbolizes features. So, the margin is calculated as $\frac{2}{||\vec{w}||}$, and to maximize the margin denominator value $||\vec{w}||$ should be minimized (Saxena, 2017).

To sum up, hard-margin support vector machine focuses on linearly separable problems whereas soft-margin support vector machine focuses on linearly separable problems with outliers.

3.5.2 Non-Linear SVM

When we are unable to separate data samples with a linear separator, non-linear SVM is applied to these data samples (Figure 3.6). This can be done by mapping input samples in a higher dimensional *feature* space. Then, since we do not want to lose advantages of linear separators, we carry out linear classification in this higher dimensional space (Figure 3.2). This is done by a kernel function.

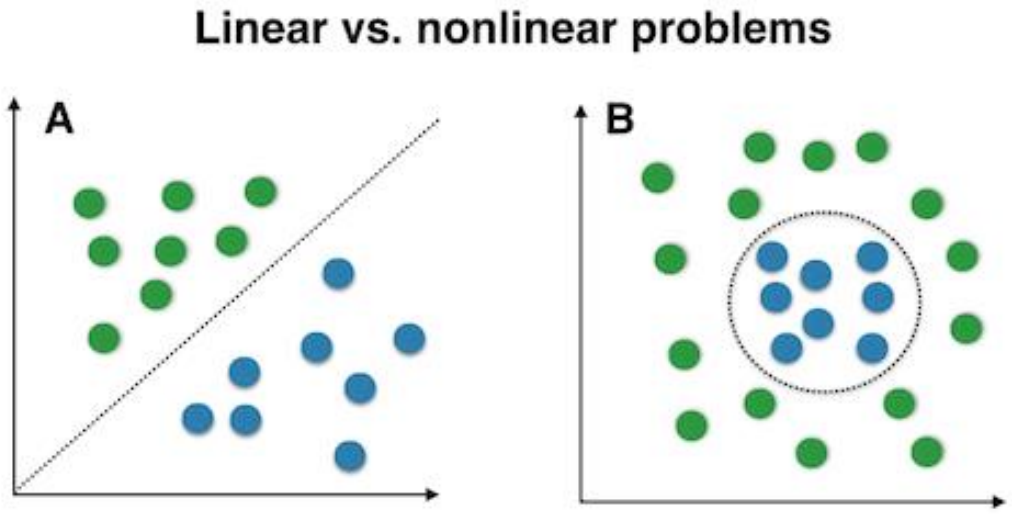

Figure 3.6 Linearly-separable and non-linearly separable problems (Chen & Bhattacharya, 2006)

Kernel functions may be declared as a dot product in a feature space (Chen & Bhattacharya, 2006). Lots of machine learning algorithms can be declared completely in terms of dot products. So, a kernel is not only restricted to support vector machines. This gives us an opportunity to replace the dot products with

kernels. For the sake of classification performance, selecting the right kernel function is important. Two of the most frequently-used kernel functions are:

- Polynomial function: $K(x, y) = (x^T y + 1)^d$ where *d* is the polynomial kernel of degree.
- Gaussian Radial Basis function: $K(x, y) = \exp(-\frac{||x-y||^2}{2\sigma^2})$ where σ represents width.

3.6 Random Forest (RF)

Random forest (RF) is a type of ensemble machine learning algorithm for classification, regression, etc., which works with multiple learning algorithms to get better prediction performance. RFs are consisting of *k* untrained Decision Trees, which are trees with only a root node and *M* bootstrap data samples. Just like in an ensemble (or in an orchestra) of instrumentalists, who play together on various instruments, when one of them plays a wrong note, the rest of the group compensate this flaw. The main idea behind ensemble learning is that a group of "weak learners" can collaborate to form a "strong learner" (Benyamin, 2012). If we obtain the mean of the results of all prediction models like voting, we can get a greater model from their combination. As the more trees are added to the forest, the forest becomes more robust which gives higher accuracy results.

Each tree grows out as far as possible, just like the over-fitting tree. However just because the formulas for constructing a single decision tree are the same every time, randomness is required to make these trees different from one another. RFs achieve this by using bagging, or bootstrap aggregating. This technique decreases model variance, or over-fitting and enhances the outcome of learning on unstable or limited number set of data samples (Lan, 2017). The working principle of bagging starts with taking the actual dataset then creating *M* subsets, where each subset contains *n* samples. The *n* samples are sampled uniformly with replacement from the actual dataset. After that, *k* ensemble (or in other words, individual learning models) are constructed for each M bootstrap sample (Lan, 2017). So, each ensembles' output is

averaged of aggregated by simple means averaging or voting (Figure 3.7). Bagging might add bias to the bagged estimator, this may cause the trade-off of reducing variance (Lan, 2017). So, this is one of things that needs to be considered.

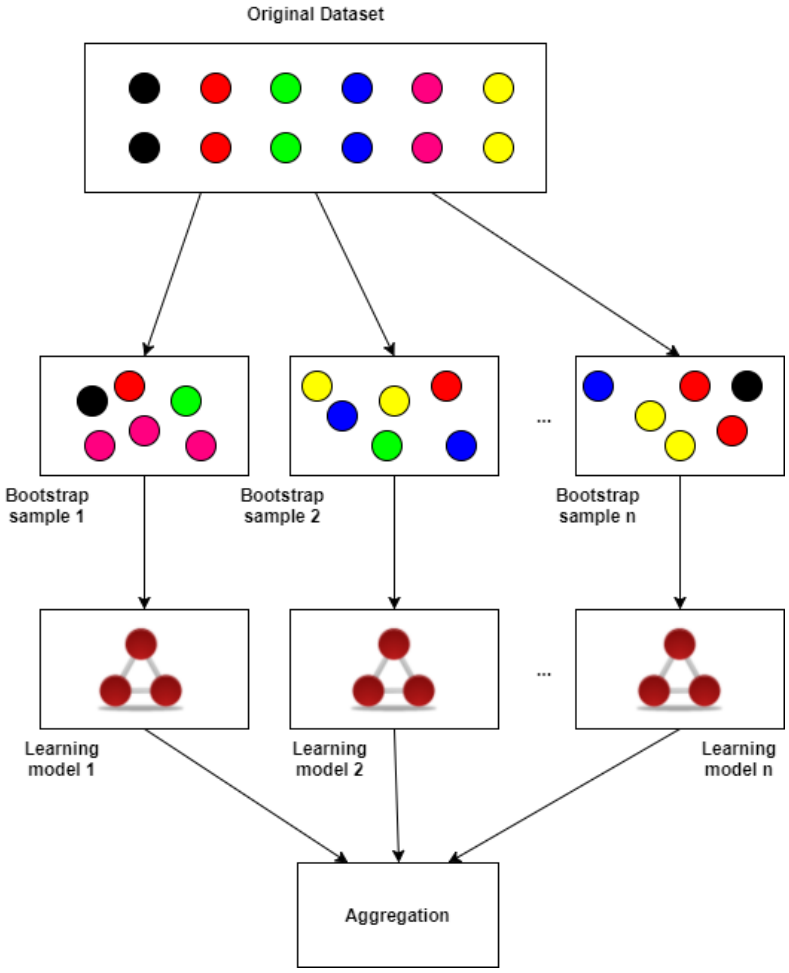

Figure 3.7 After bootstrapping and creating learning ensembles, the learning models are aggregated or averaged (Breiman, 2015)

After each tree is built by using a different bootstrap sample from the original dataset. Approximately one-third of the data are left out of the bootstrap sample and not used in the creation of the *k*th tree. This is called out-of-bag (oob) data and it is used for getting a running unbiased measure of the classification error as trees are added to the forest. That way, in RFs we do not need a separate test set or cross-validation for obtaining an unbiased estimation of the test set error (Breiman, 2015). Random forest algorithm works as follows:

1. Randomly pick *k* features at the current node from available *m* features where *k*<<*m*.
2. Then the best split point for tree *k* is calculated by using splitting metric such as information gain, or gini importance, etc.).
3. We repeat steps 1 to 2 until maximum tree depth *l* has been reached.
4. In the end, we redo steps 1 to 3 for each tree *k* in the forest and then average or vote on the output of each tree.

A random forest grows as a classification tree without any pruning. To classify a new data from an input vector, we put the input vector down each of the trees in the forest. A classification result has provided us by each tree; the class receives the highest votes between all the trees in the forest is chosen as the final class of the input vector. Decision trees may have come across with over-fitting problem. So, a procedure called pruning is applied. But, RFs accommodate a lot of trees grown with any pruning and allow them to vote for the outcome, as a result over-fitting problem does not take place.

In summary, RFs does not over-fit. We can run as many trees as we wish. RFs are fast, and they can run effectively on large data sets (Breiman, 2015). RFs provides estimates of what variables are significant in the classification. They have an efficient method for calculating missing data and maintains accuracy even when there is big part of data are missing (Breiman, 2015). RFs estimates proximities amongst pairs of cases which can be come in handy for locating outliers, or clustering (Breiman, 2015).

In this research, "randomForest" and "predict" functions of "randomForest" R-library were used. "ntree" parameter of "randomForest" function displays the number of trees to grow. It was set to 2000. Considering we own two classes, total class amount *m* is set to 2.

3.7 Individual Predictors

Two independent prediction models were constructed obtaining the same data sets, to make unbiased comparisons. The individual predictors only depended on either RPPA or RNA-seq data. The most significant *N*-genes were chosen with the t-test by comparing the sort and long survival classes, for the individual classifier of only RNA-seq data. Later, their gene expression was provided as the input of the machine learning model. The same process was done for the individual classifier of the RPPA data. Theoretically, the classification accuracy of these models ought to be lower than that of the proposed method, RPBioNet.

3.8 Cross-Validation

Prediction effectiveness of each training model is tested on an unseen (i.e., test) data sample, after then a prediction performance, such as accuracy, is calculated (Equation 3.5). Monte-Carlo cross-validation in which the whole data set was split into two sets randomly; one with 70% of the samples was set as the training set, and one with 30% was set as the test set is applied. The personalized PageRank was run with the protein level data, and the most important *N*-genes were chosen during the feature selection. SVM and RF classification models were trained with the gene expression values of these *N*-genes. The prediction effectiveness of the classification models was computed on the testing data, and the performance was calculated by the accuracy independently for SVM and RF. This train-test process was rerun 500 times with randomly splitted training and testing samples. The complete performance of each machine learning method was showed as the average accuracy of 500 iterations. In the end, the best predicting *N*-biomarkers (genes) were chosen from the individual iterations that had the topmost accuracy scores. Table 3.2 shows "predicted labels" and "actual classes" in a confusion matrix, which also known as error matrix.

Table 3.2 The confusion matrix for a two classes problem

		Real Labels	
		Class 1	**Class 2**
Predicted Labels	**Class 1**	True Positive (TP)	False Positive (FP)
	Class 2	False Negative (FN)	True Negative (TN)

The accuracy is calculated by the Equation 3.5.

$$Accuracy = \frac{(TP + TN)}{(TP+FP+FN+TN)} \tag{3.5}$$

TP represents the number of predictions (their real class is also defined as Class 1). *TN* represents the sum of predictions which are predicted as Class 2, and their real label is also Class 2.Total predictions are produced by the addition of P and N predictions. The classifiers' performance was calculated by using the receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) measures. The x- and y-axes of a ROC curve represent the false positive rate and the true positive rate, respectively. The area under the ROC curve is presented as the AUC. If the AUC value of a classifier is smaller than 0.5 for a two-class problem, this classifier performs worse than a random classifier, which turns out to be futile for a classification problem. In this research, "ROCR" and "cvAUC" R-libraries were used for those computations.

Another classification metrics are precision and recall. Precision is calculated by dividing the number of true positives (TP) by the addition of number of true positives (TP) and false positives (FP) (Equation 3.6). In other words, the number of positive predictions divided by the total number of positive class values predicted. For instance, in a medical prognostic test, precision gives us statistical information about what percentage of patients we diagnosed as having Glioblastoma actually had Glioblastoma.

$$Precision = \frac{TP}{(TP+FP)} \tag{3.6}$$

Recall (also called sensitivity) is calculated by dividing the number of true positives (TP) by addition of the number of true positives (TP) and the number of false negatives (FN) (Equation 3.7). In other words, recall gives the number of positive predictions divided by the number of positive class values in the test set. The recall answers the following question: of all the patients that actually survived, how many did we identified correctly by the test?

$$Recall = \frac{TP}{(TP+FN)} \tag{3.7}$$

3.9 Network-Based Functional Analysis

Functional enrichment analysis was practiced uncovering biological functions of the best predictive genes (i.e., biomarkers) for each cancer type supplied by RPBioNet. DAVID tool was used for the GO enrichment analysis ("DAVID Functional Annotation Tools", 2017). The ToppGene Suite was applied for both Gene Ontology (GO) and pathway enrichment analysis (Chen, Bardes, Aronow & Jegga, 2009). A network-based analysis was applied to comprehend functions of predictive genes regarding a PPI network topology.

The STRING network was applied for this purpose. The most predictive biomarkers (i.e., best achieving features out of 500 CV iterations) were merged by taking into account various cross-validation iterations whose performance are sufficient (e.g., greater than average accuracy). The runs with accuracies higher by one standard deviation of the average accuracy of each model are chosen. The number of features turned into fairly high, after choosing the CV iterations with the high accuracies. But, many of the genes were marked only once or twice in the feature sets, and despite there was a demand for filtering. If the occurrence frequency of a biomarker gene was greater than average frequency of whole biomarkers in the top performing runs, then this gene was selected as a frequent biomarker.

The biological analysis carried on by concentrating on these finest predictive and frequent biomarkers. A core network, that consist only of direct interactions amongst

the finest predictive biomarkers, was derived for each cancer on the STRING network. Afterwards, the cancer type-specific GO annotations were added into the core networks to comprehend both network-based and functional relations amongst genes. Besides, fold change (Gene expression changes) was mapped on the core network visualization. The fold change was computed by comparing mRNA measurements of patients in short- and long-survival classes. For instance, up-regulation shows greater expression in the long survival data samples.

CHAPTER FOUR
RESULTS

The main intention of RPBioNet is to perform the assignment of survival classes (short- or long term) of GBM and KIRC patients based on data retrieved from TCGA Project. mRNA expression, protein level and protein-protein interaction data were integrated to decrease the noise in experimental data and to include causal relations among the proteins. Finally, the biological interpretation of the frequent biomarker genes was found by PPI and functional enrichment analysis.

4.1 Survival Time Prediction Performance

Survival time prediction is a significant issue when it comes to creation of more personalized therapies for patients with cancer. RPBioNet divides cancer patients into classes as long- or short-term based on their survival time by making use of RNA-seq and RPPA data gathered from TCGA Project. RPBioNet uses the personalized PageRank algorithm to the human STRING Protein-Protein Interaction network to reveal the most predictive *N*-proteins based on the input of RPPA data. The impact of post-transcriptional modifications may be more precisely observed by means of RPPA data; so, the feature selection method obtains the protein measurements as the input instead of mRNA measurements. Then, the mRNA measurements of the most predictive *N*-proteins were utilized for training each machine learning method. Monte-carlo cross validation technique was applied and repeated 500 times to evaluate the learning method. RPBioNet was evaluated on 243 and 35 patients for the KIRC and GBM data sets, respectively. When the personalized PageRank algorithm was executed on two data sets, the most optimal results are gathered with N=20 features/proteins for KIRC and GBM. In Table 4.1 and Table 4.2, the average accuracy percentages of all models after applying 500 iterations of the Monte-Carlo cross-validation method is shown for GBM and KIRC, respectively.

Table 4.1 The performance percentages of all methods for GBM data set in terms of the average accuracy. The individual classifiers are trained by using either RPPA or RNA-Seq data. According to the results, RPBioNet, exceeds other methods for GBM

GBM	RPBioNet		Only RNA-Seq		Only RPPA	
	SVM	RF	SVM	RF	SVM	RF
Average-Accuracy (%)	73.3	78.3	66.8	66.7	65.0	60.2

Table 4.2 The performance percentages of all methods for KIRC data set in terms of the average accuracy. The individual classifiers are trained by using either RPPA or RNA-Seq data. According to the results, RPBioNet, exceeds other methods for KIRC

KIRC	RPBioNet		Only RNA-Seq		Only RPPA	
	SVM	RF	SVM	RF	SVM	RF
Average-Accuracy (%)	76.6	75.1	72.5	72.2	70.5	72.1

The standard deviations in accuracies while running 500 iterations for GBM and KIRC data sets are shown in Figure 4.1 and Figure 4.2, respectively. Figure 4.1 illustrates bar plot with error bars for GBM according to RPBioNet, RNASeq and RPPA. Figure 4.2 shows a bar plot with error bars for KIRC according to RPBioNet, RNASeq and RPPA. Error bars show the standard deviation of accuracies obtained by 500 iterations. The proposed method, RPBioNet, correctly predicts the survival classes with the average accuracy of 73% and 77% for GBM and KIRC patients, respectively. RPBioNet performs significantly better than the individual classifier, which is trained with either mRNA (66% for GBM, 72% for KIRC) or RPPA (65% for GBM, 70% for KIRC) data. Thus, the integration of two types of patients' data with PPI information leads the better results for the survival time prediction. For GBM, RPBioNet accomplished an average AUC of 0.6 and 0.69 by SVM and RF, respectively. Nevertheless, the average AUC values were restricted by 0.48 and 0.57 for RPPA and RNA-Seq individual classifiers, respectively. RPBioNet achieved an average AUC of 0.72 and 0.71 by SVM and RF, respectively for KIRC patients.

The approximate AUC values were calculated at roughly 0.66 and 0.69 for RPPA and RNA-Seq individual classifiers in KIRC. Both evaluation criteria, AUC and accuracy, illustrate a better performance for RPBioNet. Therefore, the integration of two types of patients' data (RPPA and RNA-seq) with PPI information provides better outcomes for the survival time prediction. SVM and RF models depict fairly similar average accuracy and AUC values for KIRC patient samples. While SVM accomplished the better accuracies in RPBioNet classifier for KIRC patient dataset, both machine learning algorithms did evenly well in terms of average AUC values.

From another point of view, RF could outrun SVM in RPBioNet classifier for GBM data samples regarding both AUC and accuracy values. If SVM and RF were compared for the GBM data set, there were 4% and 9% dissimilarity in the average AUC and accuracy values, respectively. This difference between RF and SVM machine learning models happened because of imbalanced patient samples between short-survival (24 patients) and long-survival (11 patients) classes in the GBM data set. The fundamental aim of the research is not to compare the performances of various machine learning methods by tuning their special parameters. Monitoring the effect of the new feature selection method in RPBioNet classifier is aimed; so, all machine learning algorithms were run with their default parameters, as explained before.

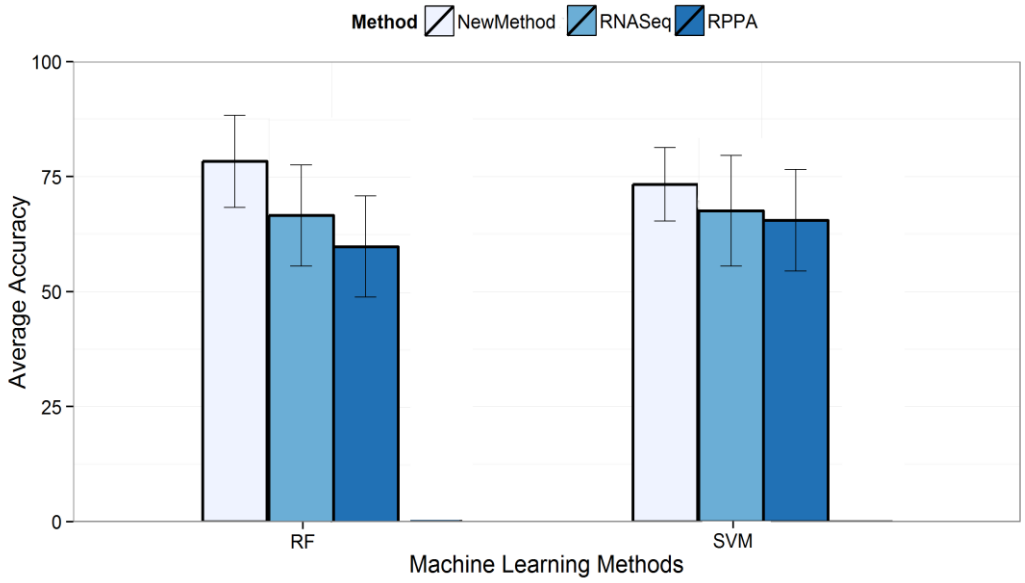

Figure 4.1 The performances of all methods in the GBM data. For SVM average accuracy measures are 0.73, 0.67, and 0.65; standard deviation results are 0.08, 0.12, and 0.11 for the new method (RPBioNet), RNASeq, and RPPA respectively. For RF average accuracy measures are 0.87, 0.66, and 0.59; standard deviation calculations are 0.1, 0.11, and 0.11 for the new method (RPBioNet), RNASeq, and RPPA respectively

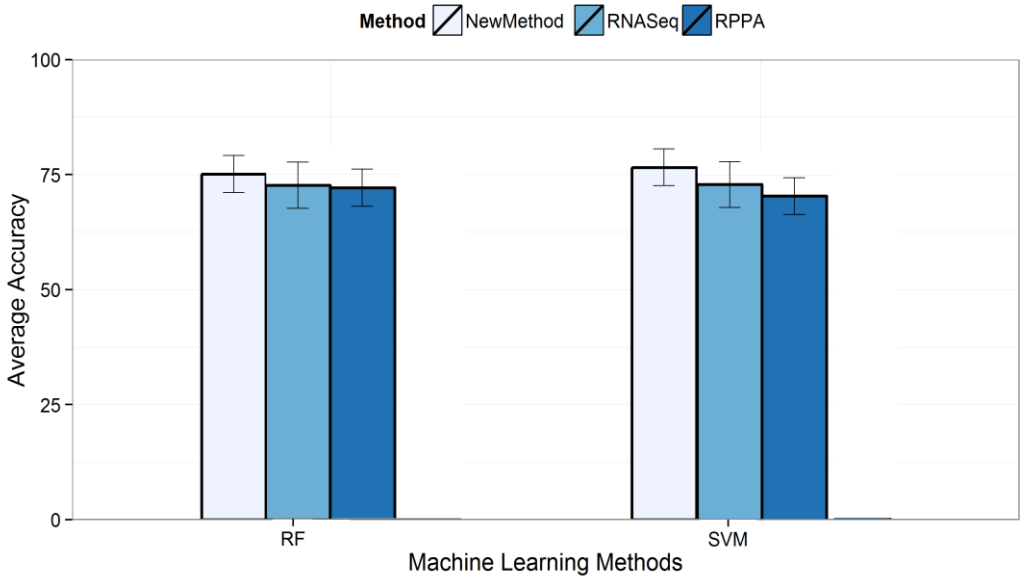

Figure 4.2 The performances of all methods in the KIRC data. For SVM average accuracy measures are 0.76, 0.73, and 0.70; standard deviation results are 0.04, 0.5, and 0.4 for the new method (RPBioNet), RNASeq, and RPPA respectively. For RF average accuracy measures are 0.75, 0.72, and 0.72; standard deviation calculations are 0.04, 0.05, and 0.04 for the new method (RPBioNet), RNASeq, and RPPA respectively

4.2 Biomarkers Biological Interpretation

After completing the run of RPBioNet on both data sets, the iterations were analyzed by looking at their accuracy percentages. The genes (features) obtaining the topmost classification accuracies should have special functions associated with progression of cancer. The most predicting biomarkers were composed by taking a few cross-validation iterations whose performance was at a satisfactory level (for instance, higher than one standard deviation of the average accuracy of each model) into consideration. It is acquired 85 and 51 different runs (for instance, different feature sets) that covered 55 and 66 unique biomarker genes for KIRC and GBM, respectively when it is applied this new filtering technique to the 500 cross-validation iterations. A few of these biomarker genes only appeared once or twice in feature sets, and so only the most frequent biomarker genes are selected. If the frequency of a biomarker gene was higher than all biomarker genes' average frequency, this gene was selected as a frequent biomarker. There were 22 and 24 frequent biomarkers for KIRC and GBM, respectively. The ToppGene Suite was used for an enrichment analysis, to comprehend the biological functions of these predictive biomarkers detected by RPBioNet. The biological annotations are detailed in Table 4.3 and Table 4.4, in which only the cancer progression-related ones are provided.

Table 4.3 The cancer-related Gene Ontology and pathway annotations for some of the most frequent biomarkers for GBM

Cancer	Annotation Type	Term	Biomarker Genes in the Term	FDR
GBM	Gene Ontology (BP)	Regulation of cell proliferation	SHC1, JUN, AKT1, KDR, ERBB3, IRS1, BCL2, CTNNA1, CTNNB1, RB1, CDH2, AR, ESR1, NOTCH1, TP53, NOTCH3, STAT3, STAT5A, SMAD4	4.34 E-13
GBM	Gene Ontology (BP)	Negative regulation of apoptotic process	JUN, AKT1, KDR, ERBB3, BCL2, UBC, CTNNA1, CTNNB1, RB1, AR, RPS6, NOTCH1, TP53, STAT3, STAT5A	3.75 E-12
GBM	Gene Ontology (BP)	Regulation of nervous system development	AKT1, KDR, BCL2, CTNNA1, CTNNB1, CDH2, NOTCH1, TP53, NOTCH3, STAT3	9.52 E-7
GBM	KEGG pathway	Neurotrophin signalling pathway	JUN, AKT1, SHC1, IRS1, BCL2, MAPK9, TP53	1.21 E-7
GBM	KEGG pathway	Glioma	SHC1, RB1, AKT1, TP53	6.77 E-3

Table 4.4 The cancer-related Gene Ontology and pathway annotations for some of the most frequent biomarkers for KIRC

Cancer	Annotation Type	Term	Biomarker Genes in the Term	FDR
KIRC	Gene Ontology (BP)	Regulation of cell proliferation	SHC1, CDKN1A, EGFR, CDH3, JUN, AKT1, AR, TP53, GAB2, CTNNA1, IGF1R, MAPK1, SRC, CCNB1, CTNNB1, IGFBP2, ERRFI1	6.13 E-13
KIRC	Gene Ontology (BP)	Negative regulation of apoptotic process	CDKN1A, EGFR, PEA15, JUN, AKT1, AR, UBC, TP53, CTNNA1, IGF1R, SRC, CTNNB1	3.79 E-9
KIRC	KEGG Pathway	Prostate cancer	CDKN1A, EGFR, AKT1, AR, TP53, IGF1R, MAPK1, CTNNB1, MAPK3	3.61 E-12
KIRC	KEGG pathway	Endometrial cancer	EGFR, AKT1, TP53, CTNNA1, MAPK1, CTNNB1, MAPK3	1.21 E-10
KIRC	PantherDB-pathway	Angiogenesis	SHC1, JUN, AKT1, MAPK1, SRC, CTNNB1, MAPK3	5.81 E-8

Some of the biological processes like cell proliferation and apoptosis, are frequently observed for two cancer types. Cell proliferation is a process that results in an increase of the amount of cells, and it is increased in tumours (Alberts, 2017). Apoptosis is a programmed cell death, and malignant cells (e.g., tumours) avoids apoptosis (Wong, 2011).

A network-based analysis was carried out to comprehend the biological relations amongst the biomarker genes at a system level, as a second step. The common (frequent) biomarker genes were mapped on the STRING network. Later, a core network, which includes only direct interactions between biomarker genes, was obtained for each cancer types. The cancer-specific GO annotations, which are

shown in Table 4.3 and Table 4.4, were also contained on top of core networks to uncover both topological and functional relations amongst biomarkers.

In Figure 4.3 two biomarker networks are shown for the KIRC and GBM data sets. The biomarker network for KIRC data set includes 9 biomarker genes, which are annotated with the "negative regulation of apoptosis" GO term and represented by an octagon shape. The network for GBM data set includes 10 biomarkers, which are annotated with the "regulation of nervous system development" GO term and shown by a rectangle shape. By comparing the expression level of patients in short- and long-survival classes, the fold change value of a gene was calculated. Therefore, a green colour represents a down-regulation in the gene expression of long-survival patient samples; orange colour represents an up-regulation in the same patient class. Nine of the biomarker genes (UBC, AR, AKT1, CTNNA1, CTNNB1, JUN, PCNA, SHC1, TP53) are common in KIRC and GBM core networks. This situation may reveal specific driver proteins linked to cancer progression.

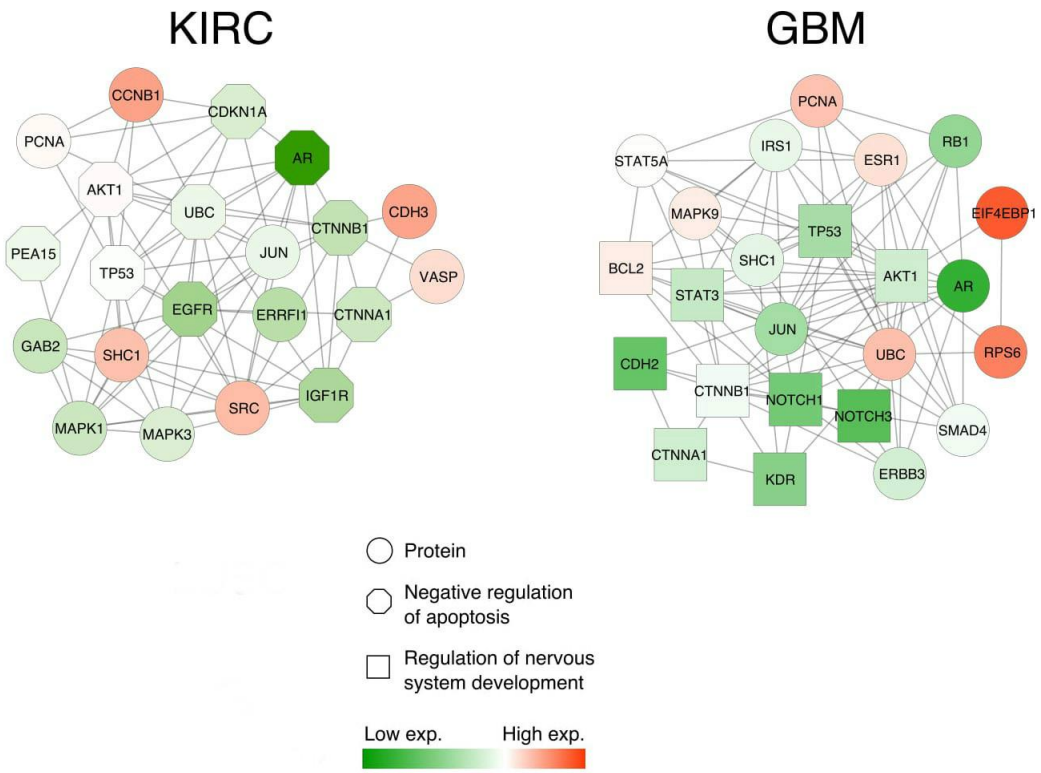

Figure 4.3 The core biomarker networks for the GBM and KIRC data sets. Circle shapes represent a biomarker gene; an edge/path between two genes shows an interaction gathered from the STRING database. In the KIRC network, each Octagon indicates the genes which are annotated with the "Negative regulation of apoptosis" GO biological process. In the GBM network, each Rectangle depicts the genes that are annotated with the "Regulation of nervous system development" process. Node color represents the fold-change value [-1.4 to 0.8] of the gene in the RNA-Seq data, for instance, higher expression signifies more mRNA measurement in the long-survival class samples or vice versa

A higher level of gene expression in biomarker genes CDH2, NOTCH1, NOTCH2, AKT1, and AR for the short-term survival patients in the GBM data sets are detected. Then a literature review is made on these biomarker genes. El Hindy et al. (2013) found out a critical mRNA increase in the GBM patients. Study by Saito et al. (2015) showed that the recurrence of GBM after radiotherapy or chemotherapy treatments in some patients with a high NOTCH1 expression. A recent study by Zhen-yi et al. (2015) discovered a positive correlation between the expression levels of NOTCH1 and EGFR genes and the GBM patients' survival time who lived more

than a year. All of these researches indicated that NOTCH1 may be a likely target for the treatment of GBM. The expression of some biomarker genes, AR (lower), SHC1 (higher), CCNB1 (higher), EGFR (lower), CDH3 (higher), and SRC (higher) were found out with miscellaneous levels in the KIRC patients' long-survival class.

During literature reviews, similar results were found with this research. According to the study by Zhu et al. (2014) the mRNA level of AR gene was discovered to be higher in normal kidney tissue. Mirza et al. (2015) aimed to find anti-cancer drugs to target S100A8 and EGFR proteins and to propose new treatment methods for kidney cancer.

CHAPTER FIVE
CONCLUSION AND FUTURE WORK

Combining multiple data sources is not a new approach for biomedical issues. Analyzing various kinds of biological networks leads to better understanding of complex cellular systems. As a result, using biological networks could have a strong influence on solving different problems such as identification of disease-causing genes, discovery of drug-targets etc.

The protein-protein interaction network that is integrated in this research has made an affirmative contribution to the identification of a limited number of biomarkers that provided more precise predictions for the future health conditions of KIRC and GBM patients. Usage of RPPA data in the PPI network for feature selection purpose is the main novelty of this study.

When the individual performances of mRNA and protein expression data are compared, the classifier trained with the mRNA data always provides better classification results than the protein data. Limited sampling of protein data in the RPPA experiments might cause this poor result, more sampling might also improve results. So, application on different (larger) data sets can improve performance of classifiers. It is also possible that optimizing parameters of SVM and RF might improve their performances. Eventually, as a future study DNA methylation data could be replacement of RPPA data in the PPI network.

REFERENCES

Adeberg, S., Bostel, T., König, L., Welzel, T., Debus, J., & Combs, S. E. (2014). A comparison of long-term survivors and short-term survivors with glioblastoma, subventricular zone involvement: a predictive factor for survival?. *Radiation Oncology*, *9*(1), 95.

Akbani, R., Ng, P. K. S., Werner, H. M., Shahmoradgoli, M., Zhang, F., Ju, Z., ... & Ling, S. (2014). A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature communications*, *5*, 3887.

Alberts, B. (2017). *Molecular biology of the cell*. Garland science.

Alexiou, G. A., Vartholomatos, E., Zagorianakou, P., & Voulgaris, S. (2014). Prognostic significance of neutrophil-to-lymphocyte ratio in glioblastoma. *Neuroimmunol Neuroinflammation*, *1*, 131-4.

Amari, S. I., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, *12*(6), 783-789.

Ananya, M. (2014) *What is a biomarker?*. Retrieved January 15, 2017 from http://www.news-medical.net/health/What-is-a-Biomarker.aspx.

Benyamin, D. (2012). *A gentle introduction to random forests, ensembles, and performance metrics in a commercial system.* Retrieved November 18, 2017, from http://blog.citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics

Beumer, I., Witteveen, A., Delahaye, L., Wehkamp, D., Snel, M., Dreezen, C., & Linn, S. (2016). Equivalence of MammaPrint array types in clinical trials and diagnostics. *Breast cancer research and treatment*, *156*(2), 279-287.

Bie, L., Zhao, G., Cheng, P., Rondeau, G., Porwollik, S., Ju, Y., et al. (2011). The accuracy of survival time prediction for patients with glioma is improved by measuring mitotic spindle checkpoint gene expression. *PloS one*, *6*(10), e25631.

Biomarker Testing. (n.d). Retrieved January 15, 2017, from https://www.nccn.org/patients/resources/life_with_cancer/treatment/biomarker_testing.aspx

Bollinger, T. (2016). *Benignand malignant tumors: what is the difference?*. Retrieved January 22, 2016, from https://thetruthaboutcancer.com/benign-malignant-tumors-difference/

Bowman, R. L., & Joyce, J. A. (2014). Therapeutic targeting of tumor-associated macrophages and microglia in glioblastoma. *Immunotherapy*, *6*(6), 663-666.

Breiman, L. (2015). Random Forests Leo Breiman and Adele Cutler. *Random Forests-Classification Description.*

Brugarolas, J. (2014). Molecular genetics of clear-cell renal cell carcinoma. *Journal of clinical oncology*, *32*(18), 1968-1976.

Buus, R., Sestak, I., Kronenwett, R., Denkert, C., Dubsky, P., Krappmann, K., & Dowsett, M. (2016). Comparison of EndoPredict and EPclin with oncotype dx recurrence score for prediction of risk of distant recurrence after endocrine therapy. *Journal of the National Cancer Institute*, *108*(11), djw149.

Chen, G. Y., & Bhattacharya, P. (2006, August). Function dot product kernels for support vector machine. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on* (Vol. 2, pp. 614-617). IEEE.

Chen, J., Bardes, E. E., Aronow, B. J., & Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic acids research*, *37*(suppl_2), W305-W311.

Choueiri, T. K., Rini, B. I., Garcia, J. A., Baz, R. C., Abou-Jawde, R. M., Thakkar, S. G., ... & Bukowski, R. M. (2007). Prognostic factors associated with long-term survival in previously untreated metastatic renal cell carcinoma. *Annals of oncology*, *18*(2), 249-255.

Christiansen, T. (2015). *Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells.* Retrieved January 27, 2017, from https://stonehilldevbio2015.wordpress.com/2015/12/02/single-cell-analysis-reveals-a-stem-cell-program-in-human-metastatic-breast-cancer-cells/

Clancy, S., & Brown, W. (2008). Translation: DNA to mRNA to protein. *Nature Education*, *1*(1), 101.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, *17*(1), 13.

Corney, D. C. (2013). RNA-seq using next generation sequencing. *Mater Methods*, *3*, 203.

Creighton, C. J., & Huang, S. (2015). Reverse phase protein arrays in signaling pathways: a data integration perspective. *Drug design, development and therapy*, *9*, 3519.

DAVID Functional Annotation Tools. (n.d.). Retrieved February 18, 2017, from https://david.ncifcrf.gov/tools.jsp

Delalat, B., Sheppard, V. C., Ghaemi, S. R., Rao, S., Prestidge, C. A., McPhee, G., et al. (2015). Targeted drug delivery using genetically engineered diatom biosilica. *Nature communications, 6*, 8791.

Edwards, B. K., Noone, A. M., Mariotto, A. B., Simard, E. P., Boscoe, F. P., Henley, S. J., ... et al. (2014). Annual Report to the Nation on the status of cancer, 1975-2010, featuring prevalence of comorbidity and impact on survival among persons with lung, colorectal, breast, or prostate cancer. *Cancer, 120*(9), 1290-1314.

El Hindy, N., Keyvani, K., Pagenstecher, A., Dammann, P., Sandalcioglu, I. E., Sure, U., & Zhu, Y. (2013). Implications of Dll4-Notch signaling activation in primary glioblastoma multiforme. *Neuro-oncology, 15*(10), 1366-1378.

Example of a RPPA workflow. (n.d.). Retrieved February 04, 2017, from https://www.fimm.fi/en/services/technology-centre/htb/instructions/workflows

Finotello, F., & Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in functional genomics, 14*(2), 130-142.

Fu, H. L., Shao, L., Wang, Q., Jia, T., Li, M., & Yang, D. P. (2013). A systematic review of p53 as a biomarker of survival in patients with osteosarcoma. *Tumor Biology, 34*(6), 3817-3821.

Glioblastoma Multiforme. (n.d.). Retrieved January 15, 2017, from http://www.aans.org/patient%20information/conditions%20and%20treatments/glioblastoma%20multiforme.aspx

Glioma. (n.d.). Retrieved January 21, 2017, from http://www.mayoclinic.org/diseases-conditions/glioma/home/ovc-20129412

Guenther, C., Hauser, A., & Huss, R. (2016). About this Book: Why Cell Therapy?. In *ADVANCES IN PHARMACEUTICAL CELL THERAPY: Principles of Cell-Based Biopharmaceuticals* (pp. 1-7).

Holland, E. C. (2000). Glioblastoma multiforme: the terminator. *Proceedings of the National Academy of Sciences*, *97*(12), 6242-6244.

Hong, B., van den Heuvel, P. J., V Prabhu, V., Zhang, S., & S El-Deiry, W. (2014). Targeting tumor suppressor p53 for cancer therapy: strategies, challenges and opportunities. *Current drug targets*, *15*(1), 80-89.

Isik, Z., & Ercan, M. E. (2017). Integration of RNA-Seq and RPPA data for survival time prediction in cancer patients. *Computers in biology and medicine*, *89*, 397-404.

Kasibhatla, S., & Tseng, B. (2003). Why target apoptosis in cancer treatment?. *Molecular cancer therapeutics*, *2*(6), 573-580.

Kobayashi, E., Ueda, Y., Matsuzaki, S., Yokoyama, T., Kimura, T., Yoshino, K., et al. (2012). Biomarkers for screening, diagnosis, and monitoring of ovarian cancer. *Cancer Epidemiology Biomarkers & Prevention*, *21*(11), 1902-1912.

Kropotov, J. D. (2016). *Functional Neuromarkers for Psychiatry: Applications for Diagnosis and Treatment*. Academic Press.

Kufe, D. W., Pollock, R. E., Weichselbaum, R. R., Bast Jr, R. C., Gansler, T. S., Holland, J. F., et al. (2003). Holland-Frei cancer medicine.

Kurzawski, G., Dymerska, D., Serrano-Fernández, P., Trubicka, J., Masojć, B., Jakubowska, A., et al. (2012). DNA and RNA analyses in detection of genetic predisposition to cancer. *Hereditary cancer in clinical practice*, *10*(1), 1.

Kush, S. (2014). *Renal cell carcinoma*. Retrieved January 15, 2017, from http://emedicine.medscape.com/article/281340-overview

Köhler, S., Bauer, S., Horn, D., & Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics, 82*(4), 949-958.

Lan, H. (2017). *Decision trees and random forests for classification and regression pt.2*. Retrieved November 19, 2017, from https://towardsdatascience.com/decision-trees-and-random-forests-for-classification-and-regression-pt-2-2b1fcd03e342

Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics, 47*(2), 106-114.

Li, B. Q., Huang, T., Liu, L., Cai, Y. D., & Chou, K. C. (2012). Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. *PloS one, 7*(4), e33393.

Loeb, K. R., & Loeb, L. A. (2000). Significance of multiple mutations in cancer. *Carcinogenesis, 21*(3), 379-385.

Luscombe, N. M., Greenbaum, D., & Gerstein, M. (2001). What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics, 1*, 83-99.

Malignant and benign brain tumors. (n.d.). Retrieved January 22, 2017, from http://www.abta.org/brain-tumor-information/diagnosis/malignant-benign-brain-tumors.html?referrer=https://www.google.com.tr/

Malinowsky, K., Wolff, C., Gündisch, S., Berg, D., & Becker, K. F. (2011). Targeted therapies in cancer-challenges and chances offered by newly developed techniques for protein analysis in clinical tissues. *J Cancer*, *2*, 26-35.

Mayeux, R. (2004). *Biomarkers: potential uses and limitations*. *NeuroRx, 1*(2), 182-188.

Mirza, Z., Schulten, H. J., Farsi, H. M., Al-Maghrabi, J. A., Gari, M. A., Chaudhary, A. G., ... & Karim, S. (2015). Molecular interaction of a kinase inhibitor midostaurin with anticancer drug targets, S100A8 and EGFR: transcriptional profiling and molecular docking study for kidney cancer therapeutics. *PLoS One, 10*(3), e0119765.

Nielsen, T., Wallden, B., Schaper, C., Ferree, S., Liu, S., Gao, D.,& Storhoff, J. (2014). Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. *BMC cancer*, *14*(1), 1.

O'Mahony, F. C., Nanda, J., Laird, A., Mullen, P., Caldwell, H., Overton, I. M., et al. (2013). The use of reverse phase protein arrays (RPPA) to explore protein expression variation within individual renal cell cancers. *JoVE (Journal of Visualized Experiments)*, (71), e50221-e50221.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.

Pe'er, D., & Hacohen, N. (2011). Principles and strategies for developing network models in cancer. *Cell*, *144*(6), 864-873.

Pucci, B., Kasten, M., & Giordano, A. (2000). Cell cycle and apoptosis. *Neoplasia*, *2*(4), 291-299.

Ray, S., Jain, K., Blog, G., & Saraswat, M. (2016). Understanding support vector machine algorithm from examples (along with code). *Analytics Vidhya*.

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

Ren, W., Li, Y., Wu, S., Feng, H., & Li, R. (2016). Protein-protein interaction (PPI) network and significant gene analysis of breast cancer. *Int J Clin Exp Med*, *9*(6), 9033-9043.

Ridker, P. M., & Silvertown, J. D. (2008). Inflammation, C-reactive protein, and atherothrombosis. *Journal of periodontology*, *79*(8S), 1544-1551.

Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B., & Peyvandi, A. A. (2014). Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterology and Hepatology from bed to bench*, *7*(1), 17.

Saito, N., Aoki, K., Hirai, N., Fujita, S., Iwama, J., Hiramoto, Y., ... & Hayashi, M. (2015). Effect of Notch expression in glioma stem cells on therapeutic response to chemo-radiotherapy in recurrent glioblastoma. *Brain tumor pathology*, *32*(3), 176-183.

Saxena, R. (2017). *SVM classifier, introduction to support vector machine algorithm*. Retrieved November 12, 2017, from http://dataaspirant.com/2017/01/13/support-vector-machine-algorithm/

Schultebraucks, L. (2017). *Support Vector Machines (SVM)*. Retrieved November 11, 2017, from https://lasseschultebraucks.com/support-vector-machines/

Shen, S., Wang, Y., Wang, C., Wu, Y. N., & Xing, Y. (2016). SURVIV for survival analysis of mRNA isoform variation. *Nature communications*, *7*.

Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics, 2016. *CA: a cancer journal for clinicians*, *66*(1), 7-30.

Souza, C. R. (2010). Kernel functions for machine learning applications. *Creative Commons Attribution-Noncommercial-Share Alike*, *3*, 29.

Statnikov, A., Wang, L., & Aliferis, C. F. (2008). A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics*, *9*(1), 319.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers?. *Current Opinion in HIV and AIDS*, *5*(6), 463.

STRING: functional protein association networks. (n.d.). Retrieved January 8, 2017, from http://string-db.org/

TCGA research network. (n.d.). Retrieved January 8, 2017, from https://gdc.cancer.gov/

Virani, S., Colacino, J. A., Kim, J. H., & Rozek, L. S. (2012). Cancer epigenetics: a brief review. *ILAR journal*, *53*(3-4), 359-369.

What is cancer?. (n.d.). Retrieved January 22, 2017, from https://www.cancer.gov/about-cancer/understanding/what-is-cancer

Wong, R. S. (2011). Apoptosis in cancer: from pathogenesis to treatment. *Journal of Experimental & Clinical Cancer Research*, *30*(1), 87.

Xing, Z. Y., Sun, L. G., & Guo, W. J. (2015). Elevated expression of Notch-1 and EGFR induced apoptosis in glioblastoma multiforme patients. *Clinical neurology and neurosurgery, 131*, 54-58.

Zhan, Y., Guo, W., Zhang, Y., Wang, Q., Xu, X. J., & Zhu, L. (2015). A five-gene signature predicts prognosis in patients with kidney renal clear cell carcinoma. *Computational and mathematical methods in medicine*, 2015.

Zhang, F., Ren, C., Lau, K. K., Zheng, Z., Lu, G., Yi, Z., et al. (2016). A network medicine approach to build a comprehensive atlas for the prognosis of human cancer. *Briefings in bioinformatics*, *17*(6), 1044-1059.

Zhang, P. W., Chen, L., Huang, T., Zhang, N., Kong, X. Y., & Cai, Y. D. (2015). Classifying ten types of major cancers based on reverse phase protein array profiles. *PloS one*, *10*(3), e0123147.

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one, 9*(1), e78644.

Zhu, G., Liang, L., Li, L., Dang, Q., Song, W., Yeh, S., ... & Chang, C. (2014). The expression and evaluation of androgen receptor in human renal cell carcinoma. *Urology, 83*(2), 510-e19.